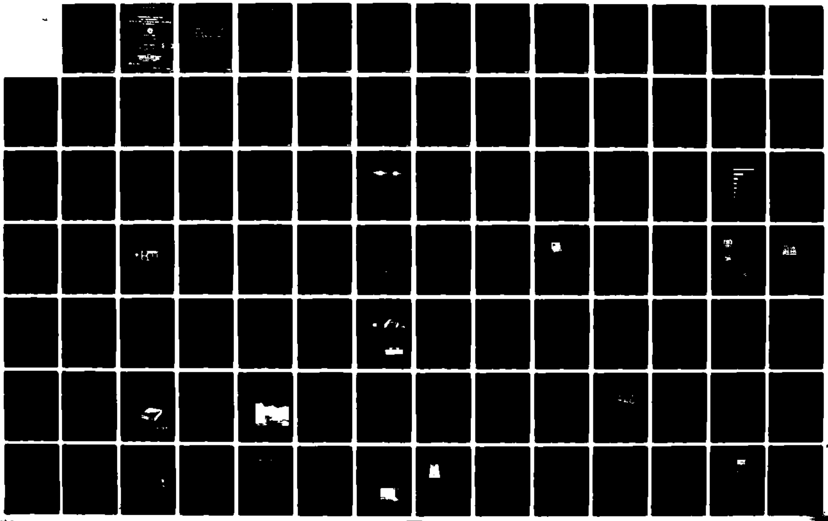
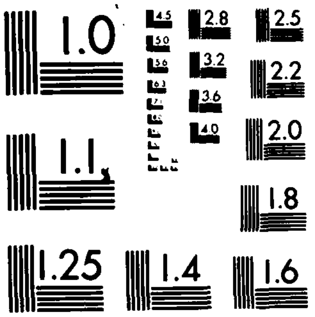
AD-A138 976 PRELIMINARY ANALYSIS OF AUTOMATIC SPEECH RECOGNITION 1/2
AND SYNTHESIS TECHNOLOGY(U) SPEECH COMMUNICATIONS
RESEARCH LAB LOS ANGELES CA J SHOUP ET AL. MAY 83
UNCLASSIFIED USCG-D-20-83 MDA904-82-C-0415 F/G 17/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Report No. CG-D-20-83

PRELIMINARY ANALYSIS OF AUTOMATIC SPEECH RECOGNITION AND SYNTHESIS TECHNOLOGY

MAY 1983

JUNE SHOUP AND JACK ROBERSON
SPEECH COMMUNICATIONS RESEARCH LABORATORY
LOS ANGELES, CALIFORNIA

Final REPORT

Document is available to the public through the
National Technical Information Service,
Springfield, Virginia 22161

[illegible]

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

The contents of this report do not necessarily reflect the official view or policy of the Coast Guard; and they do not constitute a standard, specification, or regulation.

This report, or portions thereof may not be used for advertising or sales promotion purposes. Citation of trade names and manufacturers does not constitute endorsement or approval of such products.

Technical Report Documentation Page

1. Report No.
CG-D-20-83

2. Government Accession No.
AD-A138 996

3. Recipient's Catalog No.

4. Title and Subtitle
Preliminary Analysis of Automatic Speech Recognition and Synthesis Technology

5. Report Date
May, 1983

6. Performing Organization Code

7. Author's)
June Shoup and Jack Roberson

8. Performing Organization Report No.

9. Performing Organization Name and Address
Speech Communications Research Laboratory, Inc.
3500 S. Figueroa Street, Suite 202
Los Angeles, CA 90007

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
MDA904-82-C-0415

12. Sponsoring Agency Name and Address
Department of Transportation
United States Coast Guard
Office of Research and Development
Washington, D.C. 20593

13. Type of Report and Period Covered
Final Report
July 1981 - December 1982

14. Sponsoring Agency Code
G-DMT-3

15. Supplementary Notes
The Contracting Officer's Technical Representative for this project was Dean Scribner. His guidance and assistance in this project are gratefully acknowledged.

16. Abstract

>The areas of automatic speech recognition and speech synthesis are examined to ascertain what possibilities may exist for implementing them in Coast Guard Communication Stations. A discussion of the state of the art in both speech recognition and speech synthesis is presented. Concepts from the disciplines are given, as well as descriptions of many commercially available devices.

We do not recommend that the Coast Guard pursue the development of a speech recognition system at this time. Several manufacturers are attempting to develop machines that will meet the Coast Guard's minimum requirements. Quite rapid progress is characteristic of the speech recognition field and suitable systems should be available in a few years.

We do recommend that the Coast Guard research the implementation of speech synthesis technology at this time for the following: 1) specific tasks, particularly weather reports, and 2) general purpose use in Communication Stations.

17. Key Words
automatic speech recognition
automatic speech synthesis
automation of communication systems
computer generated speech
voice recognition

18. Distribution Statement
Document is available to the United States public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
173

22. Price

Form DOT F 1700.7 (8-72)

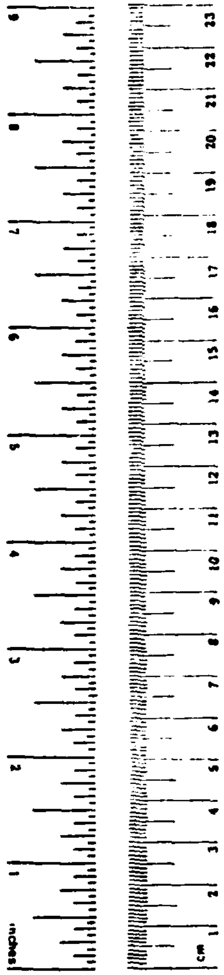
Reproduction of completed page authorized

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in ≡ 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F

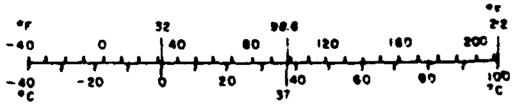


TABLE OF CONTENTS

CHAPTER page

CHAPTER 1

EXECUTIVE SUMMARY*

1.1 GENERAL EVALUATION

Commercially available recognizers may not meet Coast Guard minimum requirements, but today's synthesizers most likely will be very useful for certain tasks, such as weather reports.

1.2 RECOMMENDATION: SPEECH RECOGNITION

We do not recommend that the Coast Guard pursue the development of a speech recognition system at this time.

RATIONALE:

Current technology does not provide equipment capable of recognizing key words, such as "mayday", as found in ordinary and expected transmissions. Several manufacturers are attempting to develop machines that will meet the Coast Guard's minimum requirements. Quite rapid progress is characteristic of the voice recognition field and suitable systems should be available in at most a few years. It is highly doubtful that a duplicate research effort funded by the Coast Guard could provide suitable voice recognition units more quickly.

* A glossary of terms used in this report is given in Appendix A.

1.3 RECOMMENDATION: SPEECH SYNTHESIS

We recommend that the Coast Guard research implementation of speech synthesis technology at this time for: 1)specific tasks, particularly weather reports, and 2) general purpose use in communication stations.

RATICNALE:

A speech synthesis system could be assembled from available products to automate weather reports specifically and to be used generally in other routine transmissions within communication stations. The speech quality will be consistent, without regional accent, and the synthesis could provide for an extensive vocabulary. Since there is a tradeoff between quality and vocabulary size, it is understood that the pronunciation will be of less than broadcast standard and have some "machine quality", but will be most adequate for Coast Guard needs. The system can be made both "user proof" and "user friendly", allowing operation by personnel with little technical training. Synthesized speech could be generated instantly from reports coming over the teletype with virtually no manpower requirements. Such a system could be integrated into future general automation.

1.4 TECHNICAL BACKGROUND: SPEECH RECOGNITION

Machines differ widely in the sophistication with which they can "recognize" or "understand" human speech. The simplest systems are capable of responding only to an exceptionally limited vocabulary. For example, some can understand only the digits zero through nine, plus a very few selected vocabulary items. These words must be spoken one at a time, by one person only, and the machine will respond correctly to that person's voice only after "training", where the person says each word over and over to give the machine some idea of what to expect. These systems are said to be "isolated word recognizers" because each word must be spoken with pauses of silence as boundaries. They are not able to handle "connected or continuous speech". These systems are also said to be "speaker dependent" because they must be trained by each speaker before they are able to recognize the words correctly.

Improvements over these simplest systems are of two separate types: first, some machines can identify the vocabulary items in their list without having to be trained and some machines can receive connected or continuous speech. The advantage of the systems which have the first improvement is that anyone can communicate with the machine immediately without having to train it. Such systems are called "speaker independent", because the machine does not need information about which individual is addressing it, and, therefore, can respond independently of such information. Most all of these systems still have to have the words presented one at a time.

The second kind of improvement has resulted in machines which have fairly extensive vocabularies that can be presented to them in a relatively normal, continuous manner.

These machines can receive spoken input as ordinary sentences, rather than one word at a time. They are referred to as machines capable of handling "connected or continuous speech". They still have to be "trained" to be able to understand any individual who is going tc communicate with them.

From the point of view of virtually all applications, it is unfortunate that no one system yet incorporates both improvements, thereby becoming both speaker independent and capable of handling connected speech. Private industry, which sees a major market for improved speech recognition systems, is attempting to solve the problems involved in producing a widely useful recognition unit.

A large part of the difficulty with producing a better system is purely technical. That is, the basic approaches used so far seem likely to continue to be fruitful, but they need refinement. The heart of the problem lies in the fact that no machine actually "understands" anything in the intuitive and rational way a human being does. The machine does not "recognize" anything, either, in the way humans do. What machines actually do is follow a set of instructions which are individually very simple, yet very numerous and assembled in complex ways. We give an oral command and say that the machine "understands" it. But what we mean is only that the action produced by the last instruction is the action a human would take upon hearing the same original oral command. To get the machine to behave correctly, all the intermediate instructions between the ccmmand and the action must be correct. It is already a complex task to assemble instructions that are known to be useful. In addition, to produce improved speech recognizers, new types of instruction must be developed and integrated into the systems.

One of the central issues in speech recognition lies in the area of "template matching". Cne way for a machine to decide whether a word that it "hears" is the same as a word that it "knows" is to make a comparison between a stored "template" and the incoming word that has just been spoken. Humans do this intuitively without being able to explain how they do it. We know exactly how a machine does it, because we must give it explicit instructions on what features to consider. When it identifies an acoustic feature in both the template and the input signal, it must be told how much of that feature must be similar, if not identical, to the template to count as "the same". When it has identified all the features as "same" or "different", it must be told what proportion of all features must score as "same" to result in the whole word being considered to be an example of the "same word".

No two humans speak exactly like one another, and no one person always says the same word in the same way every time. It is difficult to give a machine instructions that allow for sufficient flexibility and at the same time preserve the essential patterns. Any number of small variations that humans handle with ease can generate wrong decisions in a machine. A word spoken more rapidly or more slowly than the template word may match at the beginning, for the first phoneme. By the time the input word arrives at the fifth phoneme, a simple machine may be trying to compare that fifth phoneme of the input with the fourth phoneme of the template. Such a comparison would result in an erroneous judgment of "different". Techniques such as time warping attempt to ensure that the machine will stretch or shrink the input to fit the template, but they have not yet reached the level of sophistication needed to assure correct matches in every case.

Analogous problems exist for input words produced at levels of loudness or stress different from the template, or words spoken with regional accents different from that of the person who produced the words on which the template is based. For these problems partial solutions exist, each roughly as successful as time warping, but none guaranteeing totally correct recognition.

A different kind of problem is presented by noise. Humans have a capacity for "selective attention" by which they automatically pay attention to the speech sounds and ignore any random hiss, crackle, bang, or other non-speech sound. As far as a machine is concerned, any sound that enters the system is as important as any other. In order to prevent the machine from trying to match irrelevant noises with features in the template, a way must be found to separate the noise from the desired signal. Methods developed to date are successful only with low levels of noise, although improvement is being made.

In short, machines are extremely different from people, and in performing tasks of speech recognition, far less competent.

Of the approximately 19 speech recognition units reviewed, not one is of a level of sophistication to meet minimum Coast Guard requirements. The rapid progress of basic research in speech recognition, however, makes it appear likely that suitable units will be available for purchase in a few years, but it is not recommended that the Coast Guard implement the technology at this time.

1.5 TECHNICAL BACKGROUND: SPEECH SYNTHESIS

Speech synthesis is the science of producing human speech by artificial means, usually by performing various operations on stored material. The stored data base may or may not ultimately derive from recorded human voices. Systems that derive from recordings have the advantage of sounding fairly natural, but they have the disadvantage of a limited vocabulary for a given set of messages and of high cost due to large storage requirements. One of the most successful of such systems is generally known as LPC synthesis. (LPC stands for "linear prediction coding", and refers to the method by which the computer selects the material it extracts from the original speech for storage, from which it will produce an imitation of the human voice.

IPC synthesis uses a digital filter to model the human vocal tract. It is based upon the statistical assumption that human speech changes relatively slowly, and that it is possible to predict the next set of acoustic measures based on a knowledge of previous ones.)

The other major class of synthesizers falls into the category of rule synthesizers (also called text-to-speech synthesizers or phoneme synthesizers) which are not based on recorded human speech. These synthesizers store a formula for the components that represent the sounds of the letters in ordinary spelling. Such sound components are called phonemes. Much as a group of letters are assembled to form a written word, so a related set of phonemes are assembled to form a spoken word. The name "rule synthesizer" emphasizes one aspect of such word-building, the fact that there are general patterns in the English language which can be described in the form of a set of rules for the computer to apply.

One set of rules involves English spelling, which is often notoriously non-phonetic. For example, "so" and "do" have the same letter at the end, but they are not phonetically pronounced the same. The rules help the computer obtain basic pronunciations for most words which follow general spelling rules. Words such as "knowledge" or "freight", where the pronunciation would be absurd if all the letters were pronounced, are treated separately. The other sets of rules, and much the more technically demanding to develop, are the ones that refine the pronunciation from the first attempt to something more acceptable to the human listener.

There are two general areas in which refinement is needed. One is that when the units of sound (individual phonemes) are first assembled, they may not blend smoothly together and form a recognizable word, even though the correct sounds are present in the right order. What is needed is a set of rules linking each phoneme to the next with suitable transitions to smooth the pronunciation and unify the sound of the word. The other area where rules are needed is to assemble the words, which may individually be acceptable, into a sentence which flows smoothly in patterns of rhythm and emphasis expected by a normal human listener.

Any sentence, however reasonable, becomes suddenly difficult to understand if the individual words are separated and said one at a time, as if in a list. What a list lacks are the subtleties of emphasis that let the listener know which words are minor parts of the utterance. This general pattern or "melody" of the sentence is usually referred to as the intonation. Developing rules for natural-sounding intonation is probably the most difficult part of text-to-speech or rule synthesis techniques, and the area in which a listener is most likely to find fault with the "machine quality" of the speech.

The advantages of rule synthesis are numerous. The storage requirements are small, making that aspect inexpensive. The vocabulary can be unlimited. Any text can be converted to speech automatically, e.g., urgent messages from the Coast Guard Communications Stations to the marine community can become speech instantly. Some pronunciations will probably be incorrect, but can be improved with respelling of the text.

Hybrid systems also exist, as do linguistically sophisticated systems requiring extensive specialized knowledge for their operation. These systems will not be discussed in detail in this report for they do not appear to meet Coast Guard Requirements.

In summary, state-of-the-art speech synthesis best lends itself to adaptation to specific Coast Guard needs, such as broadcasting weather reports, and to general use in certain applications within communication stations.

Approximately 32 synthesizers of different types available for purchase are reviewed. Of these, perhaps two or three are potentially adaptable to Coast Guard needs, although none is an exact match to Coast Guard specifications. Considerable work would be required to achieve the best possible mix of such factors as naturalness and intelligibility of speech, ease of operation, limitation of cost, and potential for integration into future large-scale communications automation. Ultimately, however, the use of synthesis for certain Coast Guard tasks, such as weather reports, would be advantageous and is recommended for consideration.

CHAPTER 2

REVIEW OF COAST GUARD'S STATEMENT OF WORK

This report discusses two advanced technologies, speech recognition and synthesis, for possible use at Coast Guard Communications Stations and Radio Stations. SCRL was asked to consider that speech recognition technology be used as an aid to watch standers for spotting distress calls over the marine VHF-FM, HF(voice), and MF(voice) frequencies. In particular, the Coast Guard has expressed an interest in keyword spotting for incoming broadcast messages (for example, automatically recognizing "mayday", "fire", "sinking", etc.). Such keyword spotting would be a means of reducing the error rate in monitoring distress frequencies. SCRL also considered that speech synthesis techniques be used for automatic broadcasting of stored text messages, such as weather information, notices to mariners, hydrographic information, storm warnings, advisories, safety messages, and urgent messages.

2.1 SPEECH RECOGNITION TECHNOLOGY AND COAST GUARD PLANNING

Two Coast Guard publications were furnished to SCRL for evaluation in this area: 1) Telecommunications Manual (COMDTINST M2000.3A), and 2) Coast Guard Radio Frequency Plan (COMDTINST M2400.1A).

The Coast Guard's Statement of Work covering this project further noted that its main area of interest in speech recognition technology involved keyword spotting in connected or continuous speech. It also stated that the Coast Guard was aware of problems involved with applying speech recognition technology to Coast Guard needs in this area. These included distortion and noise in incoming signals due to radio transmission and problems in recognition due to coarticulations in different phonetic contexts.

Perhaps the most serious problem is that the Coast Guard obviously requires a completely speaker independent recognizer. Currently available speaker independent recognizers only operate reliably with digits, that is, words such as one, two, three, four, etc. Recognition of words other than digits requires training the recognizer for individual speakers' vocabularies. Usually this is accomplished by the recognizer system prompting the user, who speaks the desired vocabulary items so they can be used as templates for matching with incoming words. Several training passes (repetitions) are usually required in order to "train" the speaker dependent recognizer. It is not feasible to obtain training material from all speakers of messages which are received by the Coast Guard. Thus, the speech recognition technology must be speaker independent to be useful for Coast Guard application.

There is no currently available speech recognizer that would meet the Coast Guard's requirements for a speaker independent recognizer capable of keyword spotting for incoming distress signals. Several companies are working diligently in this area, but it should be at least several more years before any manufacturer is able to market such a recognition system.

A second requirement for the spotting of keywords in distress signals concerns the need for a recognizer that can handle connected or continous speech. Most all voice recognizers are isolated word recognizers which require pauses between words (or simple phrases which are treated as words) so that problems of coarticulations in different phonetic contexts will be avoided. Recently, however, certain companies have marketed recognizers which are capable of handlng connected speech up to 180 words per minute. Unfortunately, these connected or continuous speech recognizers are not speaker independent. So, they do not meet both needs of the Coast Guard.

Another major technical problem with spotting keywords in Coast Guard distress signals would be that such signals have a relatively low signal-to-noise ratio. SCPL's analysis of Coast Guard signals revealed a signal-to-noise ratio which ranged from approximately 13dB up to an approximate 30dB, with an average of 23dB. Such a low signal-to-noise ratio creates a real problem for currently available recognizers. While it is true that recognizers will generally operate with a relatively high degree of background noise, it should also be noted that this background noise must be of a periodic nature or it will create false recognitions. Incoming signals, with their pops, clicks, and other nonperiodic sounds, would present problems in this area. Also, incoming distress signals received by the Coast Guard involve different speaker rates, dialects, and accents which include phonetic and prosodic variability. These areas all involve problems for speech recognition techniques and should be better resolved before the Coast Guard select any form of speech recognition technology.

2.2 SPEECH SYNTHESIS TECHNOLOGY AND COAST GUARD PLANNING

The Coast Guard may be interested in speech synthesis technology as it relates to automatic broadcasting of stored text messages, such as weather information, notices to mariners, hydrographic information, storm warnings, advisories, safety messages, and urgent messages.

One important Coast Guard consideration relating to the potential use of speech synthesis technology is that it would help to ease manpower requirements for the production of required Coast Guard broadcasts. Synthesized utterances can be readily obtained for transmission. Synthesized utterances should approach those of a trained broadcaster in overall quality. Synthesized broadcasts would have no variances in voice characteristics due to changes in the distance between the microphone and the mouth, or due to regional dialect differences. It might be added, however, that if changes in phonetic or prosodic codings are desired to produce different alertness responses, these variations can be generated by voice synthesis. Also any text can be converted to speech automatically, e.g. incoming weather reports coming over a teletype can become speech instantly. The Coast Guard also feels that digital storage techniques are, of course, more reliable and easier to maintain and edit than analog tape recordings or magnetic drums.

Currently the Coast Guard uses analog speech recordings which have good quality, but they present the following problems: a) vocabularies are limited in size, b) the recordings are difficult to modify, c) the recordings are difficult to operate automatically, and d) the recordings produce a discontinuous dialog when spliced together.

Some of the above problems might be eliminated by the use of speech synthesis. This technology produces sounds associated with basic units of speech (phonemes) which are combined to make words. Electronic logic reads stored text, assembles phonemes into words or sentences in the proper sequence and outputs the desired synthesized utterances. Prosodic characteristics (such as stress, pitch, and duration) can be modified as required to produce the desired pronunciation characteristics of synthesized utterances.

The desire to have utterances of an unlimited vocabulary and simultaneously of good quality presents a challenge in that a tradeoff exists between vocatulary size and quality. Quite obviously, a very limited number of vocabulary items (such as the digits) can be carefully synthesized to obtain very good quality. On the other hand, it is difficult to synthesize an exceedingly large vocabulary (such as 10,000 words) with an equivalently good quality. The type of synthesis used in each case would be different, as discussed below.

2.2.1 TYPES OF SPEECH SYNTHESIZERS. It is important that information be provided to the Coast Guard regarding the possible implementation of speech synthesis to meet its broadcast requirements. The various types of synthesizers available will be briefly detailed here, since different types of speech synthesizers have specific advantages and disadvantages as they relate to Coast Guard needs. Basically, there are three main types of speech synthesis:

1) Analysis synthesis or LPC synthesis - such synthesizers typically rely upon stored linear prediction coefficients which are used to define a digital filter which simulates the human vocal tract. Such synthesizers typically exhibit high-quality speech output, with realistic prosodic features, such as stress, intonation, etc. LPC synthesizers are not generally geared to the production of specific

phonemes, but are set to output whole words based upon real human speech which has been analyzed. A limitation of analysis synthesis synthesizers is that they typically require large amounts of computer storage for individual words, since words are stored as complete units.

2) Rule synthesis - such synthesizers do not rely upon an actual analysis of speech as a basis for output of synthesized utterances. Instead, rule synthesizers use combinations of different parameters which are designed to simulate actual speech. Rule synthesizers generate combinations of basic phonemes, so they typically exhibit large vocabularies. There are certain limitations to rule synthesizers. Their output speech is typically not of the same quality as LPC synthesis. Rule synthesizers have difficulty with prosodics such as stress, intonation, etc. They also encounter problems with coarticulations, since different phoneme combinations have different coarticulations.

3) Digital recordings for synthesis - devices of this type are not speech synthesizers in the literal sense of the term. The approach is to digitally record human speech which is typically stored on LSI chips for subsequent playback. One advantage to digital recordings is that they exhibit high quality audio output, since they consist of actual "recordings" of

human speech. On the other hand, such an approach does not allow one to synthesize novel utterances, or to combine phonemes to achieve a very large output vocabulary.

SCRL notes that several manufacturers of synthesizers now market, or are planning to market, text-to-speech systems which are advanced rule synthesizers. These are designed to allow the user to type in a sequence of words at a computer terminal, which are subsequently output as whole spoken sentences. Such text-to-speech synthesizers generally will include not only segmental (phoneme or letter) encoding, but also suprasegmental (prosodic) information, such as appropriate stress levels and intonation patterns, to improve their output. This is a definite advantage where the user wants to output whole sentences or phrases with natural-sounding prosodic patterns.

2.2.2 COAST GUARD CONSIDERATIONS. For all types of synthesizers, the size of the broadcast vocabulary is a primary consideration. The Coast Guard appears to require a synthesizer with a very large, if not unlimited, vocabulary. Rule synthesizers have a definite advantage in this area. A major consideration regarding the use of voice synthesis by the Coast Guard involves the degree to which vocabulary changes would have to be made. It appears that the Coast Guard would require very frequent changes in their broadcast vocabulary. Distress, safety, and urgent messages might require such changes. As noted, the rule synthesizers do have a very definite advantage in this area, as they can string phonemes together to create new words without difficulty. On the other hand, analysis synthesis typically does not include this capability, but does allow the user to string different combinations of words together, often within the context of some basic sentence to preserve natural sentence intonation. Weather broadcasts might be synthesized using this technique until more natural sounding speech is generated by rule synthesis, assuming the main vocabulary is relatively fixed. New items, such as names of storms, could be added to the fixed vocabulary as needed.

Another very important consideration involves the quality of broadcast messages. It can be assumed that the Coast Guard requires high-quality speech output, with good prosodic characteristics and no serious audio degradation of broadcast messages due to difficulties with coarticulations between phonemes. Note that analysis synthesis, such as provided in LPC synthesizers, does exhibit high-quality phonetic and prosodic characteristics, since it simulates actual human performance in this area. Rule synthesizers typically include at least several levels of stress, which must be manipulated by the user to ensure realistic-sounding output. It is important that the Coast Guard have the opportunity to evaluate the acceptability of output from different types of speech synthesizers for its use.

2.2.3 LEVELS OF SYNTHESIS PRODUCTS. It should be noted that there are several levels of synthesis products which are commercially available, just as there are several types of commercial synthesizers. These products are detailed to assist the Coast Guard to become more familiar with various options regarding implementation of voice synthesis techniques. There are basically three levels of synthesis products available. First, there are the complete systems, which come with a host computer and which require minimal software. Second, there are the board-level products, which are designed to plug into a host computer. Finally, there are the LSI chip level products which must be integrated into circuit boards before they can be used.

There are several synthesizers which come with a host computer and all relevant software. The main advantage to such synthesizer systems is that they require virtually no installation or host software. For example, Centigram markets a speech development system, complete with a digitizer, host computer, disk, and a parametric waveform synthesizer. The main advantage to such a system is that it requires no software for integration with a host computer. Such a system is particularly applicable where users might have a need for an additional host computer or do not yet have a computer system.

Board-level speech synthesizers are generally designed to plug into RS 232-C interfaces, the most common type of computer interface. Such synthesizers are available in a wide variety of configurations: analysis synthesis types, rule synthesis types, and digital recording types. These do require that the user have a host computer to control the synthesizer, as well as in many cases to store additional vocabulary which is to be synthesized. In most cases, the necessary software is supplied by the manufacturer. Note that board-level synthesizers are generally quite reasonable in price (from approximately $500 to $3,000), depending upon the synthesizer configuration desired.

Finally, speech synthesizers are available as LSI chip level products, which have to be integrated into circuit boards before they can be used. Actually, such chip level synthesizers are generally sold to original equipment manufacturers (O.E.M.) for use in consumer products, home computers, etc. Chip level synthesizers are also available in a wide variety of configurations. Integrating chip level synthesizers into a large-scale speech synthesis strategy requires a relatively high degree of engineering and electronics sophistication on the part of the user. Naturally, along with actual LSI speech synthesis chips, the user must include additional chips for storage of vocabulary, clock timer circuitry, etc. All in all, this would have to be considered the most complex approach the Coast Guard could take with regard to speech synthesis, and one that should be approached with caution. This is particulary true in that manufacturers typically do not provide necessary controller software with their LSI chip level products.

One point which should be made is that most manufacturers of speech synthesis products are very willing to work with users in the actual setup of their synthesizers. In many cases, too, manufacturers offer custom boards designed for quite specific purposes. Should the Coast Guard purchase a speech synthesizer system, it can be assumed that manufacturers would be willing to work with them closely to get their system operational. Also, manufacturers typically offer custom vocabularies for their board level products, to suit quite specific needs. This is particularly true for analysis synthesis (LPC) synthesizers which generally are not designed to string phonemes together to create new vocabulary items. The Coast Guard should be aware that custom synthesizers do require considerable lag time for the manufacturer to prepare LSI chips with desired vocabulary items and to integrate these into actual circuit boards. Changes in the vocabulary of analysis synthesis type synthesizers typically require installation of additional LSI chips containing the proper vocabulary.

2.2.4 ASSUMPTIONS REGARDING COAST GUARD SPEECH SYNTHESIS NEEDS. Based upon general considerations regarding the actual use of speech synthesizers, SCRL is able to make several important assumptions regarding the potential use of speech synthesizers by the Coast Guard. These assumptions should help clarify the type of synthesizer the Coast Guard might wish to use for meeting its broadcast requirements.

1) The use of speech synthesis strategy would avoid several of the problems the Coast Guard now faces in meeting its broadcast requirements. For example, speech synthesizers do not typically require the use of soundproof booths as do current Coast Guard broadcasts. Synthesizers also avoid the problem of speakers enunciating broadcasts at different repetition rates, or with different dialects. Speech synthesizers also avoid problems with varying distances between the mouth and microphone, inherent in analog-type recordings fo. broadcast.

2) The Coast Guard apparently requires an essentially unlimited vocabulary for its total broadcast requirements. This assumption is based upon the fact that broadcasts would have to name ships, storms, etc. If broadcasts were based upon a stored set of vocabulary items, it seems most likely that this vocabulary would have to undergo very frequent changes. This all argues for a rule synthesizer, which would have the capability to concatenate phonemes to create new vocabulary items as desired.

3) It can be assumed that the Coast Guard would require very high-quality speech synthesis, since broadcast

messages are transmitted over radio channels which would subject them to further acoustic degradation. Towards this end, the Coast Guard should have the opportunity to evaluate the output from various types of synthesizers to assure that such output would meet its needs. The Coast Guard should, if possible, check the quality of synthesized broadcasts to see if they are acoustically acceptable after they have been broadcast over Coast Guard radio channels. This would help to ensure their overall acceptability for Coast Guard purposes.

4) Whatever speech synthesis strategy the Coast Guard adapts, it should not require elaborate operator training so as to save manpower.

5) Any synthesizer system considered should be very time-efficient, having the capability to output desired broadcasts instantaneously from teletype messages without long turnaround times.

6) Any synthesizer system considered by the Coast Guard should be highly reliable, with a backup system, if possible. Actually, since currently available synthesizers rely upon LSI circuitry, they are typically very reliable for they contain no moving parts.

7) The Coast Guard should investigate fully just what software will be required for any synthesizer system it

might wish to consider and see that this synthesizer would fit in with its overall system requirements, including host computer interfacing, programming languages, etc. All types of speech synthesizers have their advantages and disadvantages. The Coast Guard should carefully weigh these before opting for any particular type of synthesizer system.

In conclusion, SCRL is optimistic about the Coast Guard's use of speech synthesis strategies in meeting its broadcast requirements. Such synthesizers should have several advantages over currently used broadcast techniques; not the least of these advantages is the relatively low price of several commercially available speech synthesizers. Finally, SCRL stresses the point that speech synthesis should present one option by which the Coast Guard should be able to save manpower in meeting its broadcast requirements, and maintain or increase the quality of Coast Guard broadcasts.

CHAPTER 3

SIGNAL ANALYSIS OF SELECTED COAST GUARD BROADCASTS

This chapter contains an acoustic evaluation of the sample Coast Guard broadcasts which were supplied to SCRL for analysis. There were 15 sample broadcasts contained on the analog tape supplied to SCRL by the Coast Guard. The tape was recorded at the U.S. Coast Guard Communications Station, Honolulu, in late 1980. The sample broadcasts were arbitrarily selected, to be typical signals received that radiomen felt represented the range of poor to excellent signals. These sample broadcasts were input to several acoustic analyses. These analyses were designed to help establish how the Coast Guard broadcasts might interact with speech input/output technologies. In particular, we wanted to determine how well Coast Guard broadcasts might interact with speech recognition technologies which are either currently available or under development.

SCRL used its Interactive Laboratory System (ILS) for analysis of the acoustic characteristics of Coast Guard broadcasts. SCRL's ILS system operates on a DEC PDP 11-45 computer, with an RSX-11-D operating system. The basic procedure for acoustic analysis of Coast Guard signals was to digitize them from analog tape and perform acoustic analyses upon these digitized waveforms. There were several steps which were carried out for the acoustic analysis of Coast Guard signals. These included:

1) Analysis of pitch and RMS amplitudes of signals. Such an analysis was carried out with in-house software. The program provided both a display of pitch contours and RMS amplitudes of input waveforms, and a hardcopy which was output by SCRL's lineprinter, as shown in Figure 1. The primary purpose of this analysis was to obtain numbers which would allow us to compute the signal-to-noise ratio for selected Coast Guard signals. Such a computation is an important measure for a preliminary analysis of how Coast Guard signals might interact with speech recognition technology, where this is a primary consideration.

2) Spectral analysis of Coast Guard broadcasts which was carried out using TLS software. Specifically, ILS was used to compute inverse filter coefficients on digitized signals. Following this, the spectral content of specific frames of input data were displayed and copied on SCRL's hardcopy unit. The purpose behind this approach was to allow for analysis of the spectral content of Coast Guard signals, as shown in Figure 2. Specifically, SCRL was interested in determining what the cutoff frequencies were for Coast Guard broadcasts, so that we would be able to arrive at a better understanding of how Coast Guard broadcasts might interact with speech recognition technology.

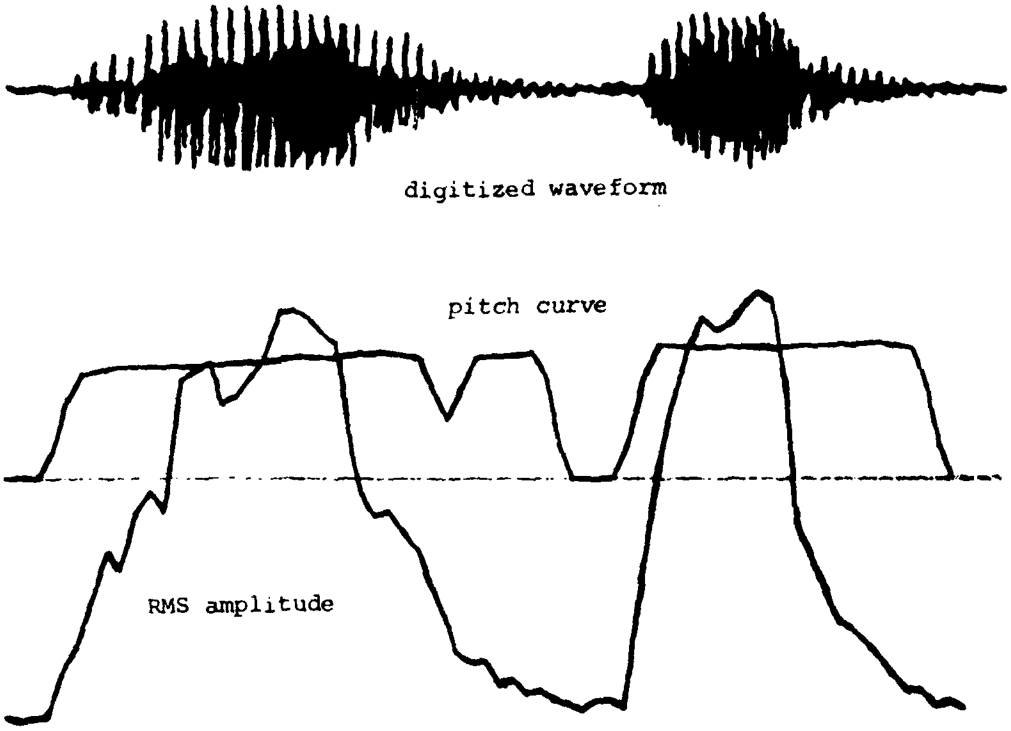


Figure 1: Sample printout from SCRL pitch and RMS detection program.

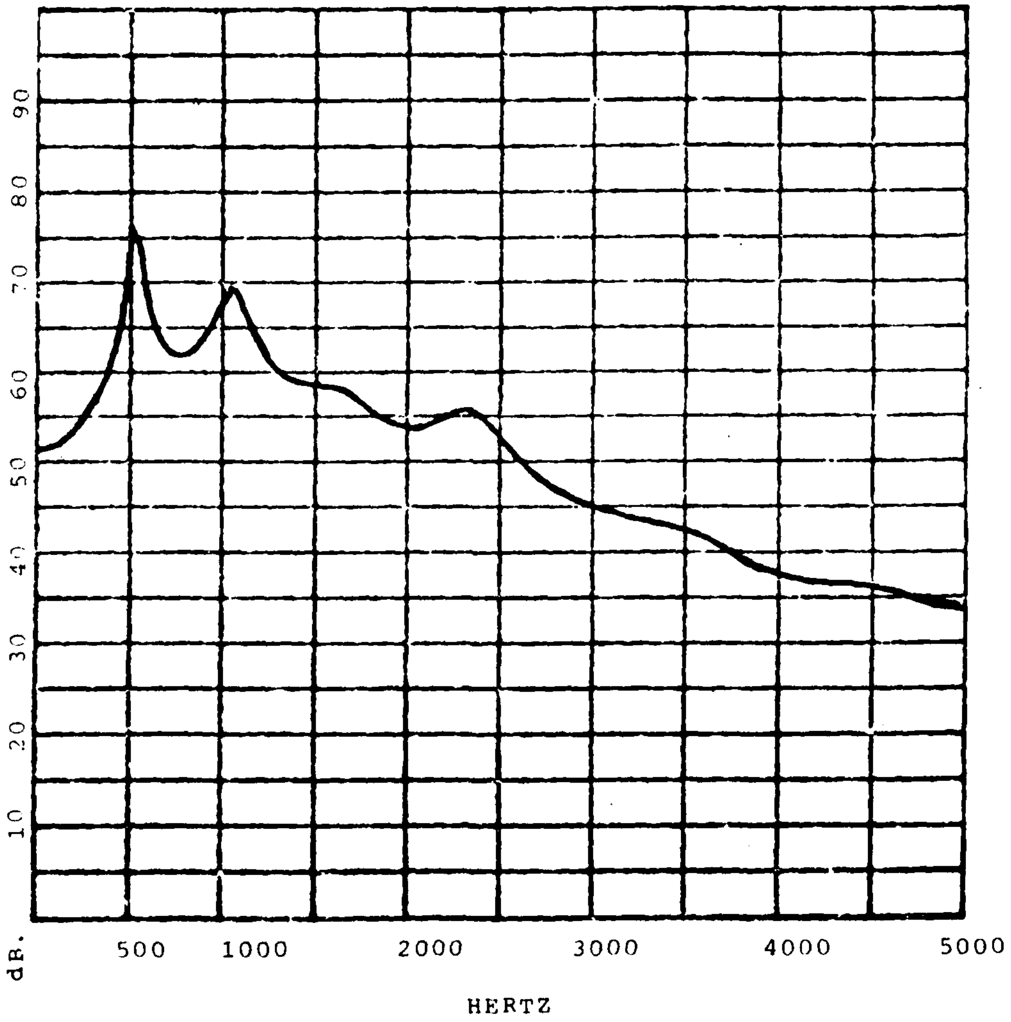


Figure 2: Spectral plot from sample Coast Guard broadcast.

Note that most speech recognizers input data from approximately 500 Hz. up to approximately 5-6 kHz. Thus, we wanted to determine whether or not Coast Guard signals were within this range, and how much background noise these broadcasts generally contained.

SCRL computed signal-to-noise ratios from RMS amplitude measurements carried out with SCRL's in-house autocorrelation pitch extraction program. Table 1 has the signal-to-noise ratios computed for sample Coast Guard broadcasts analyzed by SCRL.

TABLE 1

Signal-to-Noise Ratios for 15 Sample Coast Guard Broadcasts

Broadcast #	Signal-to-Noise Ratio	Approximate Cutoff Frequencies
1.	21.87	400-4000 Hz.
2.	13.12	300-3500 Hz.
3.	18.00	500-4000 Hz.
4.	24.74	400-4000 Hz.
5.	26.84	300-3500 Hz.
6.	25.75	600-3000 Hz.
7.	24.90	400-3500 Hz.
8.	28.46	400-3500 Hz.
9.	30.69	500-3500 Hz.
10.	29.29	700-4000 Hz.
11.	15.80	700-3500 Hz.
12.	23.72	500-3500 Hz.
13.	18.15	500-4000 Hz.
14.	20.10	500-4000 Hz.
15.	21.26	400-3500 Hz.
Mean	22.85	
Variance	5.13	

Descriptive statistics were performed on this corpus of data. It was noted that the 15 sample Coast Guard broadcasts examined had a mean signal-to-noise ratio of 23dB., with a standard deviation of 5dB. With regard to cutoff frequencies, it was noticed that Coast Guard broadcasts fell generally within the 300-4000 Hz.range. Approximate cutoff frequencies were used to compute signal bandwidths. The term "bandwidth" is generally used to describe signals which are confined to a distinct region of the frequency spectrum. Bandwidth is a useful term in the present context, since it can be used to describe the width of Coast Guard signals, in terms of their Hertz range. Sample Coast Guard broadcasts evidenced a mean bandwidth of 3207 Hertz, with a standard deviation of 315 Hertz.

In view of the poor signal-to-noise ratio which was inherent in the Coast Guard recordings, it was not practical to do further acoustic analysis of the speech signal. It can be assumed that identification of phonetic and prosodic details in these recordings would be difficult for most analysis techniques. The poor signal-to-noise ratio inherent in Coast Guard messages received would drastically limit the use of speech recognition systems for potential applications, such as wordspotting. The transmission of synthetic speech should be well tested to be certain acceptable quality is maintained regardless of the signal-to-noise ratio.

CHAPTER 4

OVERVIEW OF SPEECH RECOGNITION PRODUCTS AND TECHNOLOGY

This chapter of the report gives an overview of the speech recognition products reviewed. Basically, SCRL notes that nearly all manufacturers listed recognition accuracy rates in the vicinity of 99% for their recognizers. Actually, these figures must be approached with some caution. Note that there is no established vocabulary which has been consistently used to compare recognizers. Potential users of a speech recognition system should decide what vocabulary they might wish to use with it and actually try this vocabulary out in the field under real test conditions.

Naturally, some vocabularies are much easier for recognizers than others, and results vary accordingly. For example, where vocabularies contain words with very similar phonemes, such as "right" and "ripe", items can be confused. Also note that different speakers may influence test results, depending on their cooperativeness in training and using the system. Background noise can affect accuracy results, particularly if the noise contains nonperiodic sounds. Variables relating to recognition accuracy rates should be identified in advance, and their potential impact upon recognition accuracy should be fully considered before actually purchasing any particular system. Most manufacturers are very cooperative in arranging demonstrations and evaluations of their speech recognition devices.

The bar graph in Figure 3 shows general prices of various recognition systems manufactured by Nippon Electric Company, Threshold Technology, Interstate Electronics, Votan, Heuristics, Centigram, Auricle, Scott Instruments, and Voicetek.

This bar graph basically includes top-of-the-line recognition systems for ease of comparison. As can be seen, prices for recognition systems vary from several hundred dollars up to approximately $33,000. A main point with respect to the bar graph of prices is that, in general, the higher the price, the larger the recognition vocabulary, and the faster words may be input to the system. Subsections of this report detail the advantages and disadvantages of various systems; these should be considered before selecting any particular system.

As was noted previously, the Coast Guard apparently requires a completely speaker independent recognizer for use in spotting keywords in distress signals of connected speech. As this report notes, there is no such system currently available. All recognizers which were evaluated require training by specific speakers. Interstate

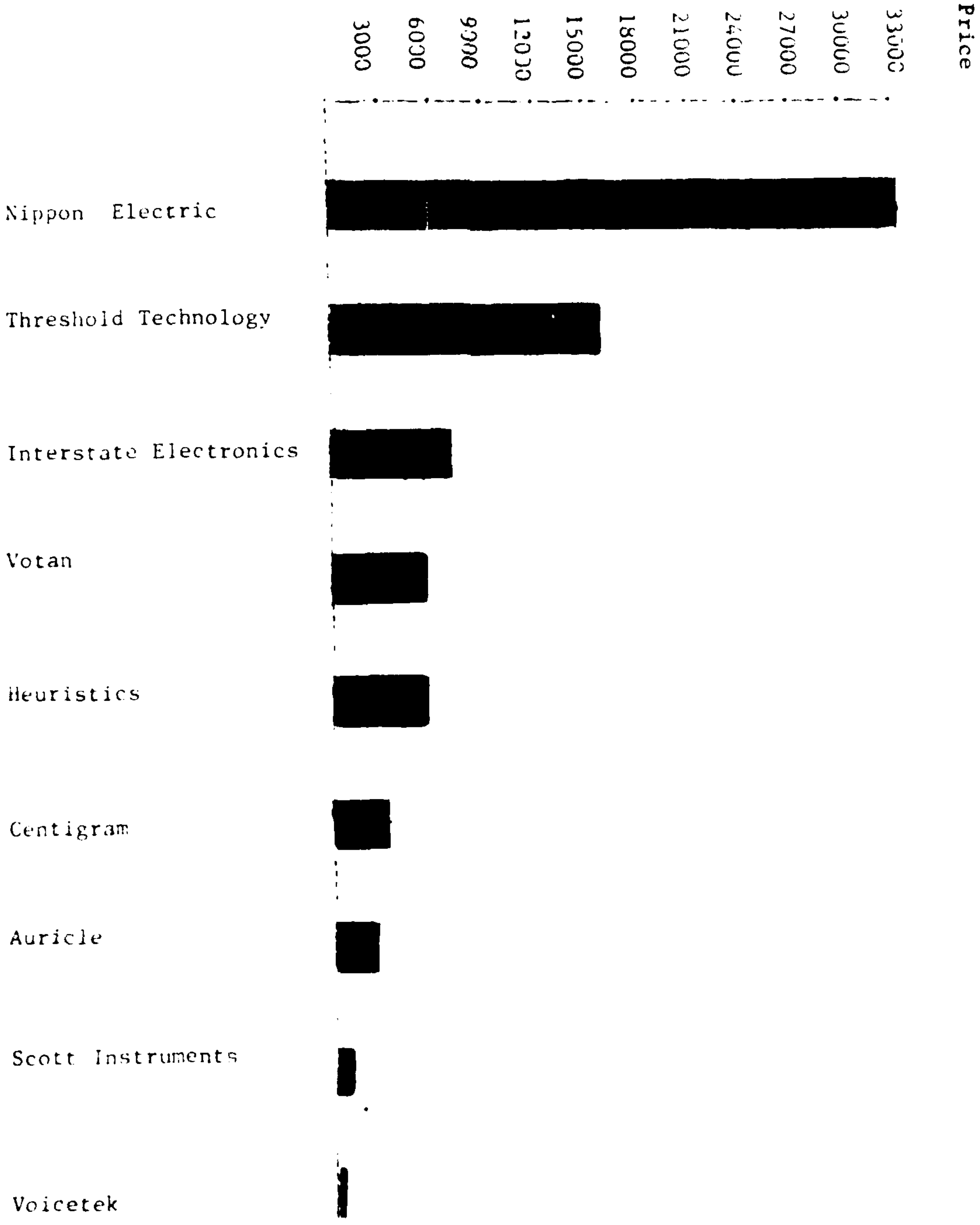


Figure 3: Recognition system prices - by manufacturers

Electronics does market a speaker independent recognition chip which has a claimed accuracy rate of approximately 85% for the general population, using digits only. We single out for Coast Guard consideration the Verbex 1800 recognizer, which handles up to 50 isolated words plus digits and control words, in a speaker-independent fashion. In fact, we suggest the possibility of using this recognizer to handle keywords which might reasonably be assumed to be somewhat isolated, or spoken at a fairly slow rate.

There is a continued effort among manufacturers to develop a completely speaker independent voice recognition system. However, it should be two or three years before anyone succeeds in marketing such a system which is capable of meeting Coast Guard needs in the area of keyword spotting in distress signals, using connected speech. There are many technical problems associated with developing such a system. Obviously, different speakers can have quite different acoustic manifestations for particular phonemes, coarticulations, overall accents, etc. Another obvious problem is the difficulty in handling speech recognition using connected speech, where words may be quite different acoustically than their isolated acoustic forms and segmentation of word boundaries is seldom clear. Nonetheless, manufacturers are competitively developing speaker-independent recognizers which can handle connected speech input. Advances will surely be made in this area within the near future. As we have stated, we expect that there may be a commercially available, speaker-independent recognizer capable of handling a relatively large vocabulary (approximately 100 words) in connected speech in approximately two or three years.

4.1 AVAILABLE SPEECH RECOGNIZERS

Data concerning speech recognizers were basically compiled during the period of 1981 and 1982 from manufacturers' specification sheets and brochures, as well as from direct input from manufacturers and published articles relating to speech input devices. Thus, much of our information came directly from information we requested from manufacturers of voice recognition systems. It should be noted that new product lines may have been introduced since the original collection of the data.

There is an increasing variety of speech recognition products being marketed by different companies for various applications. Price and performance parameters of these different devices vary considerably depending on the level of the products: chip level, board level, or system level. They also vary depending on the type of recognizer: isolated word recognizer vs. connected speech recognizer and speaker-dependent recognizer vs. speaker-independent recognizer. The information given below is from nine manufacturers of speech recognizers: 1)Centigram, 2)Heuristics, 3)Interstate Electronics, 4)Nippon Electric

Company, 5) Scott Instruments, 6) Threshold Technology, 7) Verbex, 8) Voicetek, and 9) Votan. The information received from each manufacturer was not always complete or of the same type of material given by other manufacturers. Consequently, it is difficult to provide comparative data for all speech recognition systems.

Centigram

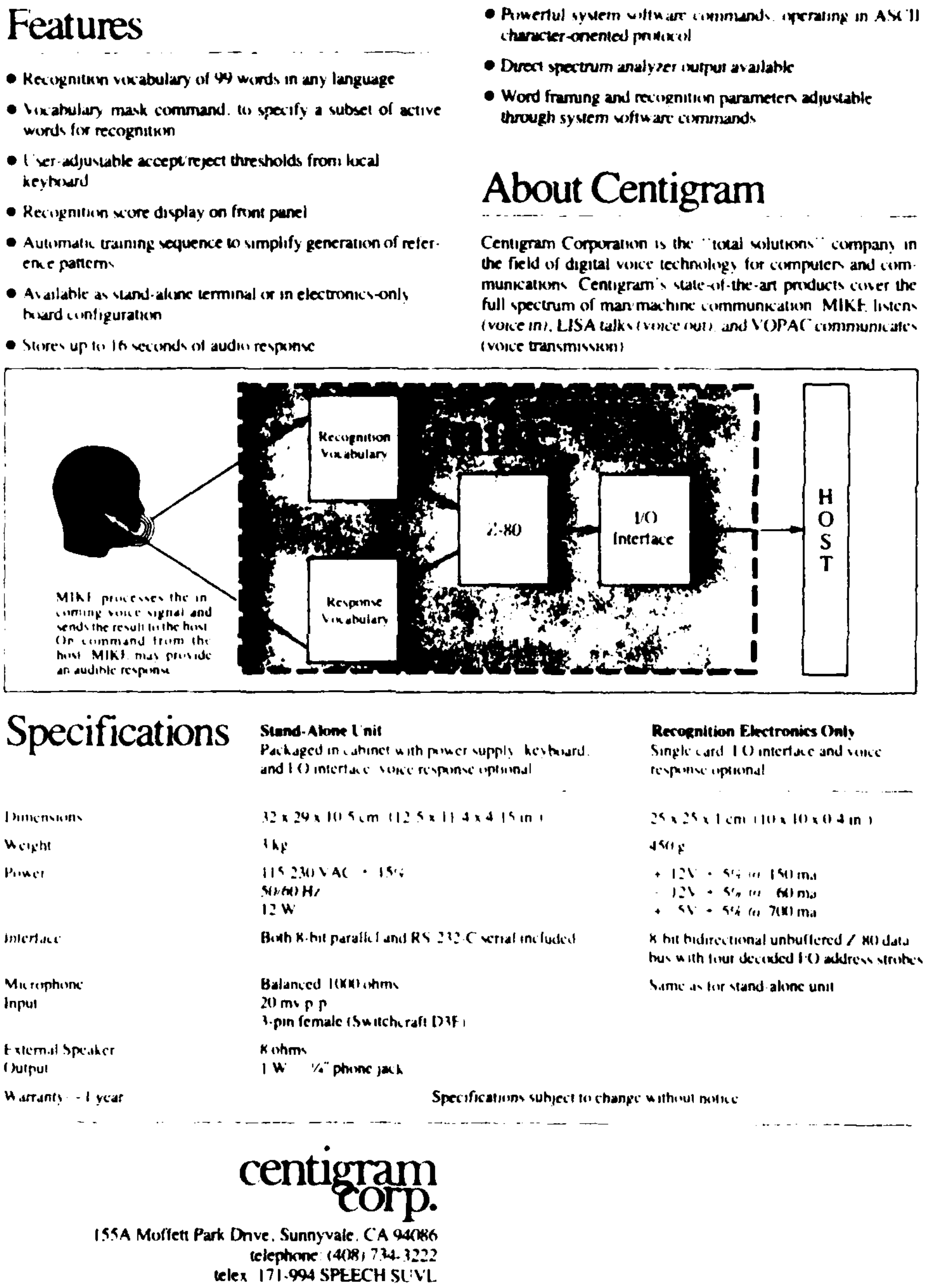
Centigram markets speech recognition products in the systems and board level categories. Their Mike recognizer is basically a speech recognition and response terminal. The Mike is noted to have two basic operating processes. Voice input is received through a recognition process for training reference patterns or performing recognition. The results of recognition are transferred to a host computer through the input/output interface. For recording audio response messages, voice input is received through a response process with the unit synthesizing recorded response messages on command from the host computer.

Table 2 lists the main features of the Centigram Mike along with operating specifications.

The Mike uses a spectrum analysis approach, where input waveform data are either stored as templates or compared to existing templates. Note that direct spectrum analyzer output is available with the Mike unit.

TABLE 2

Centigram Mike



Features

- Recognition vocabulary of 99 words in any language
- Vocabulary mask command, to specify a subset of active words for recognition
- User-adjustable accept/reject thresholds from local keyboard
- Recognition score display on front panel
- Automatic training sequence to simplify generation of reference patterns
- Available as stand-alone terminal or in electronics-only board configuration
- Stores up to 16 seconds of audio response
- Powerful system software commands, operating in ASCII character-oriented protocol
- Direct spectrum analyzer output available
- Word framing and recognition parameters adjustable through system software commands

About Centigram

Centigram Corporation is the "total solutions" company in the field of digital voice technology for computers and communications. Centigram's state-of-the-art products cover the full spectrum of man/machine communication. MIKE listens (*voice in*), LISA talks (*voice out*), and VOPAC communicates (voice transmission).

MIKE processes the incoming voice signal and sends the result to the host. On command from the host, MIKE may provide an audible response.

Specifications

	Stand-Alone Unit Packaged in cabinet with power supply, keyboard, and I/O interface; voice response optional	**Recognition Electronics Only** Single card, I/O interface and voice response optional
Dimensions	32 x 29 x 10.5 cm (12.5 x 11.4 x 4.15 in.)	25 x 25 x 1 cm (10 x 10 x 0.4 in.)
Weight	3 kg	450 g
Power	115/230 VAC ± 15% 50/60 Hz 12 W	+ 12V ± 5% @ 150 ma − 12V ± 5% @ 60 ma + 5V ± 5% @ 700 ma
Interface	Both 8-bit parallel and RS-232-C serial included	8-bit bidirectional unbuffered Z-80 data bus with four decoded I/O address strobes
Microphone Input	Balanced 1000 ohms 20 mv p-p 3-pin female (Switchcraft D3F)	Same as for stand-alone unit
External Speaker Output	8 ohms 1 W — ¼" phone jack	

Warranty — 1 year

Specifications subject to change without notice

centigram corp.

155A Moffett Park Drive, Sunnyvale, CA 94086
telephone: (408) 734-3222
telex: 171-994 SPEECH SUVL

Recognition and voice response are two different procedures, or sets of algorithms, as the Mike unit will provide up to 16 seconds of audio response, but will provide a recognition vocabulary of 99 words (in any language). Centigram also notes that recognition electronics are available on a single card, with voice response being an option. Also available is the complete stand-alone unit, with voice response also an option.

The Mike terminal sells for $4,765; the unit includes the Mike terminal, head microphone, and recognition support software. The Mike recognizer can be used in conjunction with Centigram's Lisa speech synthesizer (described in the speech synthesis products section).

Heuristics

Heuristics markets speech recognition products which are in two main areas: systems level products (terminals) and board level products. Heuristics' terminal products are moderately priced, ranging from approximately $4,600 to $5,000.

SCRL received from Heuristics descriptive brochures for two main speech recognition systems. First, Heuristics markets the 5000 Series products line. This product line is headed by the 5600 Voice Terminal System. This system features a Lear-Siegler terminal, a voice controller board with a 128-word vocabulary, disk drive and cables, and a noise-cancelling microphone. The 5600 system has a list price of $4,995. Second, Heuristics markets the 7000 Series products line. This system does not include a terminal. It consists of a voice controller board (also with a 128-word vocabulary), disk drive and cables, and a noise-cancelling microphone. It has a list price of $4,595. Note that these two systems may be purchased in part or in whole, as needed.

The Heuristics 5000 Series is described as a stand alone intelligent data entry device. The unit utilizes a spectrum analyzer using digital filtering techniques to analyze audio input and to convert it to a digital representation. Heuristics states that proprietary algorithms transform the data into a reference template. Reference templates, obtained during training, are compared with those from audio input. When a match occurs, a user-defined ASCII string is sent to the host and/or the terminal.

The following is a list of features of the 5000 Series recognizer:

1) 128 word/phrase vocabulary

2) each word/phrase may be up to 3 seconds in duration

3) 127 user-definable vocabulary subsets

4) ASCII strings up to 255 characters in length

5) local storage eliminates the need for host programming

6) simultaneous use of voice and key entry

7) device listens continuously through wraparound speech buffer

8) compatible with several languages, including FORTRAN, COBOL, PASCAL, and BASIC

9) automatic self test and fault isolation

10) RS 232C (20mA current loop) serial interface

11) 50 - 9600 Baud

12) high level auxiliary input for telephone or tape recorder (1 VRMS)

13) single board unit easily installed in Lear Siegler terminals

Table 3 provides a listing of Heuristics operation specifications for the Series 5000.

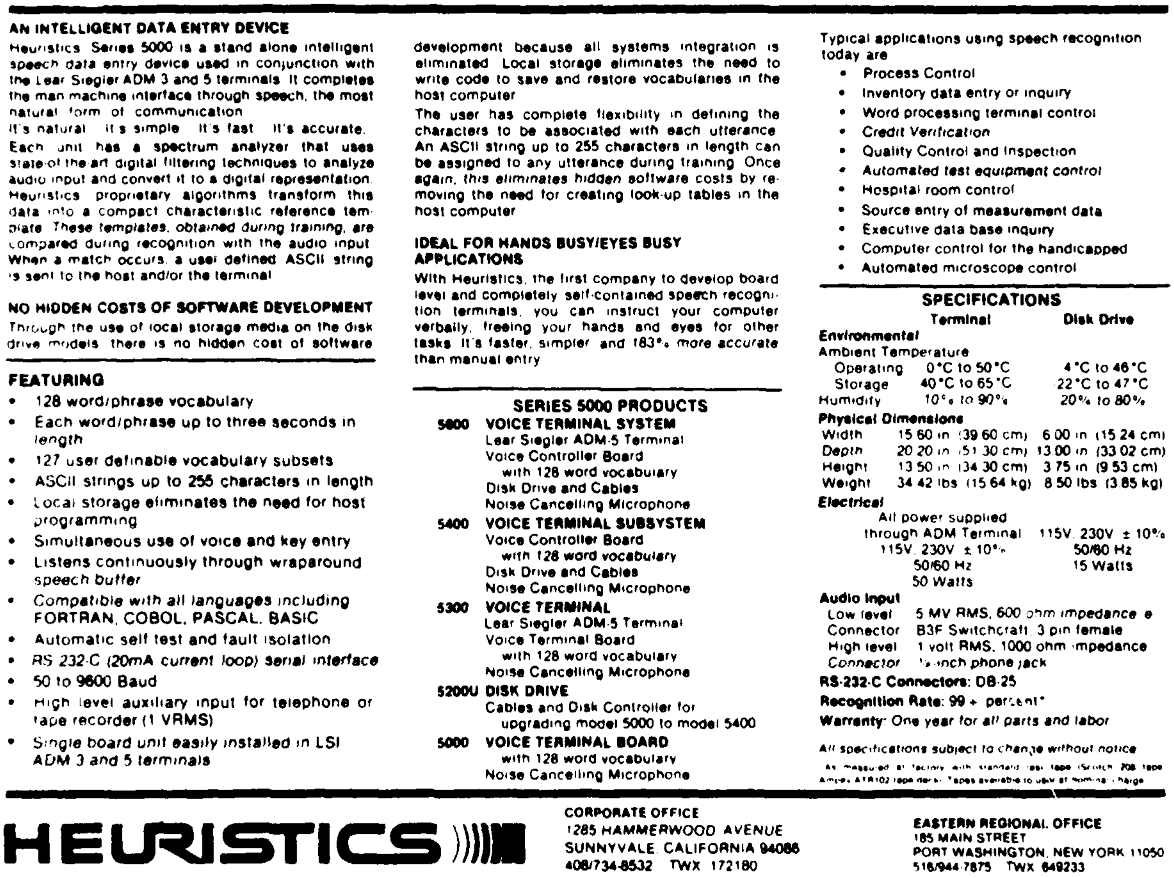
The Heuristics Series 7000 speech recognizer is similar to the Series 5000 device, but does not include a computer terminal. It is designed for use in conjunction with any ASCII terminal. This recognizer utilizes the same basic recognition approach as the Heuristics 5000 Series units. It also shares the same operation specifications with the 5000 Series units, as listed in Table 3.

Interstate Electronics

Interstate Electronics is one of the oldest companies involved in speech recognition, and currently manufactures a rather wide array of speech recognizers and related products.

TABLE 3

Heuristics Series 5000



AN INTELLIGENT DATA ENTRY DEVICE

Heuristics Series 5000 is a stand alone intelligent speech data entry device used in conjunction with the Lear Siegler ADM 3 and 5 terminals It completes the man machine interface through speech, the most natural form of communication

It's natural It's simple It's fast It's accurate.

Each unit has a spectrum analyzer that uses state-of-the-art digital filtering techniques to analyze audio input and convert it to a digital representation. Heuristics proprietary algorithms transform this data into a compact characteristic reference template *These templates, obtained during training, are* compared during recognition with the audio input When a match occurs, a user defined ASCII string is sent to the host and/or the terminal

NO HIDDEN COSTS OF SOFTWARE DEVELOPMENT

Through the use of local storage media on the disk drive models there is no hidden cost of software development because all systems integration is eliminated Local storage eliminates the need to write code to save and restore vocabularies in the host computer

The user has complete flexibility in defining the characters to be associated with each utterance An ASCII string up to 255 characters in length can be assigned to any utterance during training Once *again, this eliminates hidden* software costs by removing the need for creating look-up tables in the host computer

FEATURING

- 128 word/phrase vocabulary
- Each word/phrase up to three seconds in *length*
- 127 user definable vocabulary subsets
- ASCII strings up to 255 characters in length
- Local storage eliminates the need for host programming
- Simultaneous use of voice and key entry
- Listens continuously through wraparound speech *buffer*
- *Compatible with all languages including* FORTRAN, COBOL, PASCAL, BASIC
- Automatic self test and fault isolation
- *RS 232-C (20mA current loop) serial interface*
- 50 to 9600 Baud
- High level auxiliary input for telephone or tape recorder (1 VRMS)
- *Single board unit easily installed in LSI ADM 3 and 5 terminals*

IDEAL FOR HANDS BUSY/EYES BUSY APPLICATIONS

With Heuristics, the first company to develop board level and completely self-contained speech recognition terminals, you can instruct your computer verbally, freeing your hands and eyes for other tasks It's faster, simpler and 183% *more accurate* than manual entry

SERIES 5000 PRODUCTS

5800	**VOICE TERMINAL SYSTEM** Lear Siegler ADM-5 Terminal Voice Controller Board with 128 word vocabulary Disk Drive and Cables Noise Cancelling Microphone
5400	**VOICE TERMINAL SUBSYSTEM** Voice Controller Board with 128 word vocabulary Disk Drive and Cables Noise Cancelling Microphone
5300	**VOICE TERMINAL** Lear Siegler ADM-5 Terminal Voice Terminal Board with 128 word vocabulary Noise Cancelling Microphone
5200U	**DISK DRIVE** Cables and Disk Controller for upgrading model 5000 to model 5400
5000	**VOICE TERMINAL BOARD** with 128 word vocabulary Noise Cancelling Microphone

Typical applications using speech recognition today are

- Process Control
- Inventory data entry or inquiry
- Word processing terminal control
- *Credit Verification*
- Quality Control and Inspection
- *Automated test equipment control*
- Hospital room control
- Source entry of measurement data
- Executive data base inquiry
- Computer control for the handicapped
- Automated microscope control

SPECIFICATIONS

	Terminal	Disk Drive
Environmental		
Ambient Temperature		
Operating	0°C to 50°C	4°C to 46°C
Storage	40°C to 65°C	22°C to 47°C
Humidity	10% to 90%	20% to 80%
Physical Dimensions		
Width	15 60 in (39 60 cm)	6 00 in (15 24 cm)
Depth	20 20 in (51 30 cm)	13.00 in (33 02 cm)
Height	13 50 in (34 30 cm)	3 75 in (9 53 cm)
Weight	34 42 lbs (15 64 kg)	8 50 lbs (3.85 kg)
Electrical		
	All power supplied through ADM Terminal	115V. 230V ± 10%
	115V. 230V ± 10%	50/60 Hz
	50/60 Hz	15 Watts
	50 Watts	

Audio Input

Low level	5 MV RMS, 600 *ohm impedance* e
Connector	B3F Switchcraft 3 pin female
High level	1 volt RMS, 1000 ohm impedance
Connector	¼ inch phone jack

RS-232-C Connectors: DB-25

Recognition Rate: 99 + percent*

Warranty: One year *for all parts and labor*

All specifications subject to change without notice

* As measured at factory with standard test tape (Scotch 208 tape Ampex ATR102 tape deck) Tapes available to user at nominal charge

HEURISTICS

CORPORATE OFFICE
1285 HAMMERWOOD AVENUE
SUNNYVALE, CALIFORNIA 94086
408/734-8532 TWX 172180

EASTERN REGIONAL OFFICE
185 MAIN STREET
PORT WASHINGTON, NEW YORK 11050
516/944-7875 TWX 649233

Interstate's speech recognition products are broadly divided into 3 areas: 1)terminal products, 2)board products, and 3)voice semiconductor products.

Interstate Electronics markets two basic voice entry terminals. One of these is the VRT101 voice recognition terminal. This unit has a 100 word recognition vocabulary, and boasts a 99%+ accuracy rate. The unit includes a Z80 microprocessor, a 48K memory, and a 100K floppy diskette drive. The VRT101 sells for $5,295, with quantity discounts available. Interstate also markets the VRT103 fully integrated voice recognition terminal, which is the same as the VRT101, but includes two additional floppy diskette drives with a 300K memory capacity. The VRT103 sells for $6,595 in single units. Following are Tables 4 and 5 covering the specifications of the VRT101 unit. It should be noted that a wide variety of options are available for Interstate's speech recognition terminals.

Interstate's board products include:

1) VRM041 voice recognition module. This unit includes a 40 word voice recognition vocabulary with 99% accuracy, and RS232-C interfacing. The VRM041 sells for $1,790 in single units.

2) The VRM102 voice recognition module. The unit features a 100 word vocabulary with 99% accuracy, and 2 serial RS232-C or 20-mA asynchronous interfaces. It sells for $2,255 in single units.

3) The VRT200 voice recognition module. The unit permits voice recognition with the popular Lear Siegler ADM 3A and ADM 5 terminals. The unit features a maximum of 100 isolated words or phrases and recognition accuracy of 99% or better. The VRT 200 also includes a user programmable reject level. The VRT200 sells for $2,100 in single units.

Tables 6 and 7 provide a listing of specifications for the Interstate VRT200 voice recognition terminal.

Table 8 indicates specifications for the Interstate VRM041 and VRM102 voice recognition modules

Interstate markets a single-board speech recognition module for DEC Q-Bus equipment, called the VRQ400 voice recognition module. This board costs $2,120. It features a 100 word recognition vocabulary, trainable for any vocabulary in any spoken language. The unit is usable with direct microphones, wireless microphones, or via telephone.

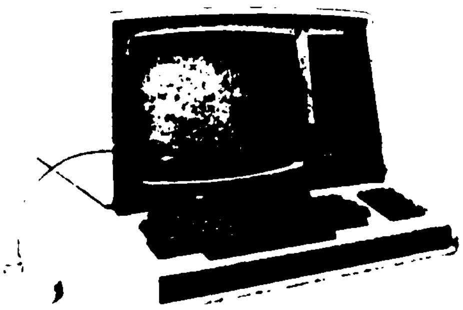
A new product from Interstate is the VRC008 voice recognition chip. This chip features an 8-word vocabulary and is speaker-independent. An accuracy rate of 85% is given for the general population. The chip has an initial charge of $25,000 for tooling and mask generation. After this initial tooling charge, the VRC008 sells for $22.50, with discounts for quantities over 25,000 units. Table 9 provides a listing of specifics regarding the VRC008.

TABLE 4

Interstate Electronics VRT101 Voice Recognition Terminal

INTERSTATE ELECTRONICS CORPORATION

VOICE RECOGNITION TERMINAL
Model VRT101



Voice recognition is fully integrated into Interstate's diskette-based VRT101 intelligent voice terminal.

- **Direct voice interaction with application software**
- **Supports a variety of application programs and higher level languages**
- **100-word resident vocabulary**
- **99%+ accuracy**
- **Utility software for immediate use of voice recognition functions**
- **Self-test for fault isolation**

MODEL VRT101 SPECIFICATIONS

CPU and Memory

Processor: Z80
Clock: 2 048 MHz
Memory: 48K bytes RAM

Display

CRT: 12 inches diagonal. P4 phosphor
Display Format: 24 lines of 80 characters plus 25th user-status line
Display Size: 6.5 inches high x 8.5 inches wide.
Character Size: 0.2 inch high x 0.1 inch wide (approximate).
Character Set: 128 (95 ANSII. plus 33 graphics).
Character Type: 5 x 7 dot matrix (upper case), 5 x 9 dot matrix (lower case with descenders).
Keyboard: 72 keys (60 alphanumeric, 12 function control) plus a 12-key numeric pad.
Cursor. Blinking or reverse video block or off.
Cursor Controls: Up, down left, right, home, CR, LF, back space, and tab from keyboard or computer.
Cursor Addressing: Relative and direct.
Tab: Standard 8-column
Refresh Rate: 60 Hz at 60 Hz, 50 Hz at 50 Hz line frequency.
Edit Functions: Insert and delete character or line.
Erase Functions: Erase line from beginning of line to end of line; erase page from beginning of page to end of page.
Bell: Audible alarm on receipt of ASCII BEL.
Video: Normal and reverse by character.

Serial Input/Output Ports (2)

Interface: EIA RS-232C at data rates of 110 to 9600 bits per second
Communication Mode: Full or half duplex.
Parity: Even odd. or none.

Disk Systems

Built-in: 5-1/4-inch floppy. 100K bytes
VRTDK2: Two external 5-1/4-inch floppies. 200K bytes.

Software

CP/M: Operating system software.
BASIC: Microsoft.
FORTRAN: Microsoft.

TABLE 5

Interstate Electronics VRT101 - Further Specifications

Mechanical

Dimensions: 13 inches high x 17 inches wide x 20 inches deep.

Weight: 54 pounds.

Environmental

Operating Temperature: 10 to 35°C

Storage Temperature: 0 to 35°C.

Power

120/240 volts at 50/60 Hz at 90 watts maximum.

Voice Recognition Performance

Vocabulary Size: 100 isolated words and/or phrases.

Recognition Accuracy: 99+ percent.

Reject Threshold: User selectable.

Longest Utterance Duration: 1.25 second.

Minimum Between-Word Pauses: 160 milliseconds (user-programmable: 40 to 320 milliseconds).

Minimum Word Length: 80 milliseconds (user-programmable: 80 to 160 milliseconds).

Processing Time: (25+N) milliseconds, where N = active vocabulary size, following detection of the end of word.

Voice Utility Commands

1. Train
2. Update
3. Reset
4. Set RTHL
5. Read RTHL
6. Download reference patterns only
7. Upload reference patterns only
8. Download reference patterns and ASCII strings (joint)
9. Recognition with common vocabulary
10. Recognition of non-contiguous vocabulary (host mode only)
11. Test (standalone mode only)
12. Write word boundary parameter
13. Read word boundary parameters
14. Set Operational Mode (standalone=0, host=1)
15. Set Gain
16. Read Gain
17. Compare Reference Patterns
18. Self-test

CP/M™ is a trademark of the Digital Research Corporation

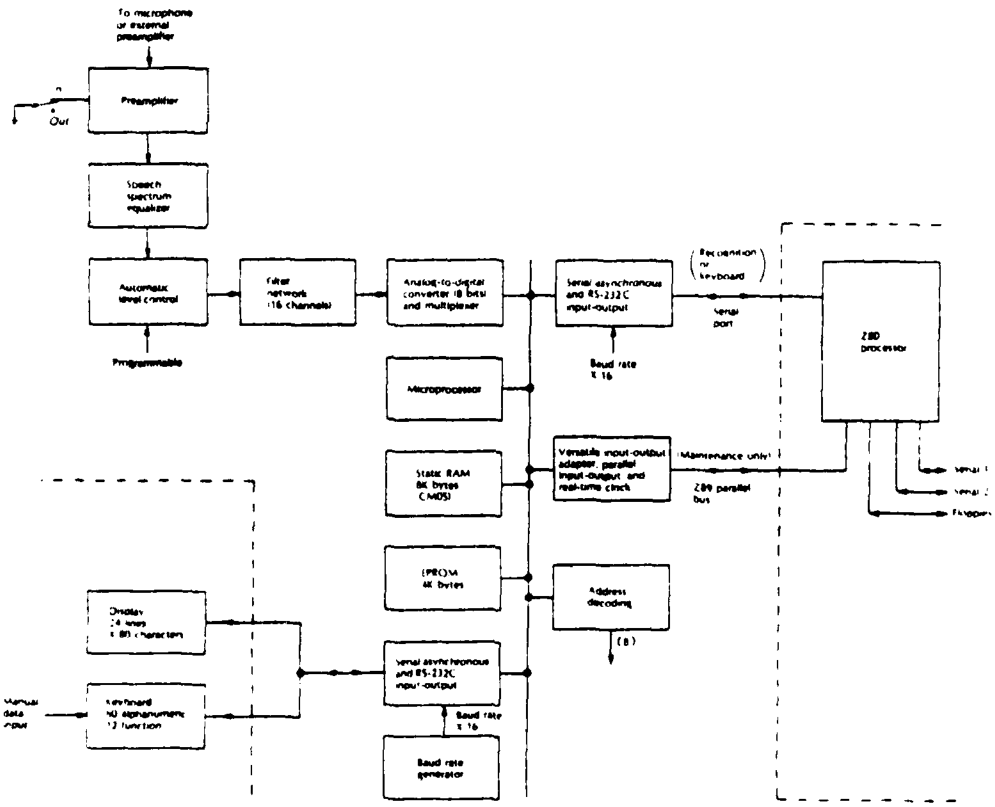
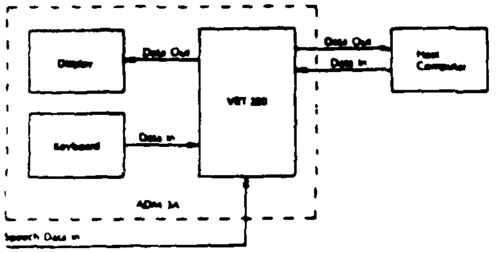

To microphone or external preamplifier
In
Out
Preamplifier
Speech spectrum equalizer
Automatic level control
Programmable
Filter network (16 channels)
Analog-to-digital converter (8 bits) and multiplexer
Microprocessor
Static RAM 8K bytes (CMOS)
EPROM 4K bytes
Serial asynchronous and RS-232C input-output
(Recognition or keyboard)
Serial port
Baud rate x 16
Versatile input-output adapter, parallel input-output and real-time clock
(Maintenance only)
Z80 parallel bus
Address decoding
(8)
Z80 processor
Serial 1
Serial 2
Floppies
Display 24 lines x 80 characters
Keyboard 40 alphanumeric 12 function
Manual data input
Serial asynchronous and RS-232C input-output
Baud rate x 16
Baud rate generator

TABLE 6

Interstate VRT200 Voice Recognition Terminal

INTERSTATE
ELECTRONICS
CORPORATION

VOICE RECOGNITION TERMINAL Model VRT200



Block Diagram of ADM 3A Terminal with VRT200 Voice Recognition Capability

- **Single-board speech recognition module**
- **Adds voice input capability to ADM 3A and ADM 5 Dumb Terminal^R Video Displays**
- **No special programming required**
- **100-word vocabulary**
- **99% + accuracy**
- **Selectable decision threshold for rejection of unwanted inputs**

Accurate, Low-Cost Automatic Speech Recognizer

The VRT200 is a single printed-circuit board speech recognizer with a vocabulary of 100 words or short phrases designed specifically for use in the Lear Siegler ADM 3A and ADM 5 Dumb Terminal^R Video Displays. With 99+ percent accuracy, the VRT200 allows Lear Siegler *Dumb Terminal*^{R} users to input commands or data via voice and/or keyboard, thus providing maximum operator efficiency for data entry, retrieval, and log-on.

The VRT200 is a total hardware/software system designed for easy installation without modification to existing application software. All VRT200 logic is contained on a single printed-circuit board that has been specifically designed to fit the ADM 3A and ADM 5 and can be installed without soldering or special tools. An ADM 3A or equivalent fitted with a VRT200 board immediately adds voice input capability to already operational data entry, process control, or management information systems.

The VRT200 allows direct microphone input via either a boom-mounted, lightweight, noise-cancelling microphone or, at the user's option, a table-mounted microphone. The microphone has a standard five-foot cord, but longer cables are available if more freedom of movement is required. Using the VRT200 frees the operator from the need to return to a fixed workstation to enter data, thus increasing operator efficiency.

Efficient, Real-Time Performance

Speech input is analyzed by a 16-channel spectrum analyzer and converted to a digital representation of the spoken input. This digital data is then converted to a fixed-size pattern that preserves the information content of the spoken inputs while discarding redundant features. During vocabulary training, these patterns are used to derive templates for each utterance. The templates are then used in the recognition process for comparison with incoming spoken words. Vocabulary templates are stored in an onboard random-access memory (RAM), while the processing algorithms are contained in an onboard read-only memory (ROM) operating in conjunction with a microprocessor. When an utterance is recognized, a user-defined ASCII string is then sent to the host.

Trainable to Individual Voice Characteristics

The VRT200 is a speaker-dependent voice recognition device, which requires that each user give a sample of the words and phrases in the vocabulary before the VRT200 will recognize the user's voice. The process of generating these samples, or reference patterns, is called vocabulary training. Once a reference pattern set has been built, it can be uploaded to the host computer's mass storage. Later, the user can download the reference patterns, allowing the terminal to recognize the same words without the need for retraining. With each VRT200, software is supplied in FORTRAN and BASIC languages demonstrating the host-resident code necessary to perform the upload/download operation.

The VRT200 supports three training modes: (1) normal training in which the vocabulary is cleared, then trained a selected number of samples; (2) updating of word patterns in which the stored reference patterns for the specified vocabulary are augmented by additional training; and (3) a single-word retrain mode in which the single word will be trained the same number of samples as the word it is replacing.

TABLE 7

Interstate VRT200 Voice Recognition Terminal - Further Details

Specifications

ADM 3A and ADM 5

CRT Screen:
12 inch (30 5 cm) diagonal P4 phosphor non-glare surface 5 8 inches (14 7cm) high x 8 3 inches (21 1 cm) wide Display 1920 characters 80 char line by 24 lines

Character Set:
ADM 5—128 ASCII characters upper lower case punctuation control characters ADM 3A—64 ASCII characters, displayed as upper case plus punctuation and control

Character Font:
Character matrix ADM 5—5 x 9 dot matrix, including full 2 dot descenders (1 88mm wide x 5 53mm high) ADM 3A—5 x 7 dot matrix (1 88mm wide x 4 77mm high)

Cursor:
Reverse block Homes to upper left of screen Optional switch selectable underline cursor homes to bottom left Switch selectable non destructive space after carriage return

Visual Attributes:
Reverse video reduced intensity and reverse video reduced intensity combination —ADM 5 only

Keyboard:
ADM 5—83 keys 26 letter alphabet with upper lower case numeric keypad punctuation caps lock cursor control All keys are auto repeating (22 char sec) ADM 3A—59 keys 26-letter alphabet with upper case numeric 0 through 9 punctuation control Two key repeat operation (22 char sec)

Transmit/Receive:
Conversation mode full half duplex

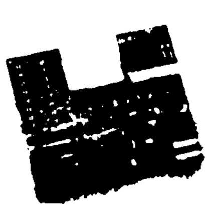
Interfaces:
RS 232C point to-point or 20mA current loop

Data Rates:
75-19 200 Parity Even odd mark space or none

Word Structure:
Data 7 or 8 bits, 1 start bit, 1 or 2 stop bits

Extension Port:
RS 232C port for interfacing serial asynchronous devices

VRT200



Vocabulary Size:
Maximum 100 isolated words or phrases

Percent Recognition Accuracy:
99 percent or better

Reject Threshold:
User programmable

Longest Utterance Duration:
1 25 seconds

Minimum Between-Word Pauses:
160 milliseconds

Minimum Word Length:
50 milliseconds

Approximate Response Time:
(25 + N) milliseconds where N = active vocabulary size (following end-of-word detection)

Software:
Provides vocabulary reference pattern upload/download for DEC RT-11 DEC RSX-11M, Data General RDOS

Standard Power Requirements:
115 Volts ± 10%, 60 Hz, 60 watts

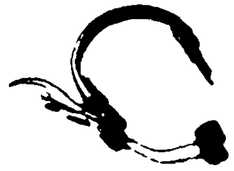
Optional Power Requirements:
230 Volts ± 10%, 50 60 Hz

Width:
12 5 inches

Height:
10 5 inches

Operating Environment:
5° to 50°C (41° to 122°F) 5% to 95% relative humidity without condensation, 10,000' (3km) max altitude

Microphones



VMK010:
Noise cancelling microphone, microphone mounted on aluminum headset includes ON/OFF switch for audio input control

VMK012:
Noise-cancelling microphone with earphone Same as VMK010 except includes earphone for voice response

VMK545:
Stand-mounted microphone Cardioid pickup pattern with gooseneck and stand including ON/OFF switch

VMK577:
Hand-held microphone For use in applications where continual voice entry is not required Contains noise-cancelling element and PUSH-TO-TALK switch

INTERSTATE ELECTRONICS CORPORATION
1001 East Ball Road
P.O. Box 3117
Anaheim, California 92803
(714) 635 7210 (800) 854-6979
TWX 910-591 1197
Telex 655443 & 655419

Regional Offices

Western Office
1001 East Ball Road
Anaheim, California 92803
(714) 635 7210

Northwestern Office
500 Airport Blvd, Suite 110
Burlingame, CA 94010
(415) 342 8624

Central Office
1375 Remington Road, Suite M
Schaumburg, IL 60195
(312) 843 7233

Eastern Office
1745 Jefferson Davis Hwy, Suite 601
Arlington VA 22202
(703) 892 1400

Lear Siegler, Inc.
Data Products Division
714 North Brookhurst Street
Anaheim, California 92803
(714) 774-1010
TWX 910-591 1157 Telex 65-5444

Boston (617) 890-7093
Chicago (312) 279 5250
Houston (713) 780 2585
Los Angeles (213) 454-9941
New York (800) 523-5253*
Orlando (305) 869 1826
Philadelphia (215) 245-1520
San Francisco (415) 828-6941
Washington D C (800) 523 5253*
England (04867) 80666

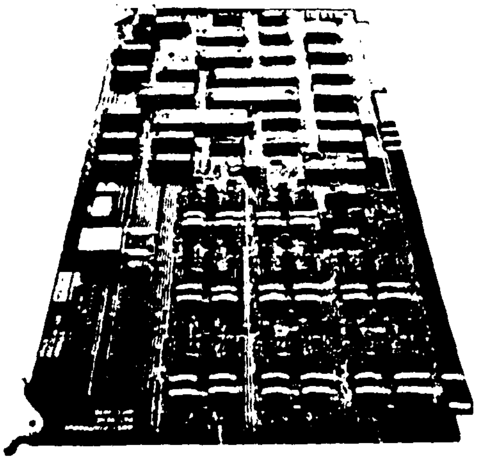
*800 number also includes CT DE MA MD NJ NY RI VA & WV

TABLE 8

Interstate VRM041 and VRM102 Voice Recognition Modules

INTERSTATE
ELECTRONICS
CORPORATION

VOICE RECOGNITION MODULE
Models VRM041 and VRM102



Interstate's single-board Voice Recognition Module

- **99%+ accuracy**
- **40- and 100-word vocabularies**
- **Highly accurate real-time operation**
- **Trainable for any vocabulary in any spoken language**
- **Multibus form factor**
- **Usable with direct microphones, wireless microphones, or via telephone**
- **User selectable rejection of poor input match**
- ***One parallel and two serial ASCII input/output ports***
- **User control of recognition parameters**

Accurate, Low-Cost Automatic Speech Recognizer

The Voice Recognition Module (VRM) is a single printed-circuit board speech recognizer capable of recognizing as many as 100 words or short phrases. It is easily interfaced to an external system using either parallel or serial interfaces. The serial interfaces are switch selectable to RS232-C or 20-mA current loop. The VRM includes all the logic and memory necessary to perform training, word recognition, and the communication protocol independent of the user's mode of operation.

The VRM contains a microphone preamplifier and a preamplifier bypass switch to allow direct *microphone* input using a lightweight headset, boom-mounted, or hand-held microphone. Alternately, an audio signal may bypass the onboard preamplifier, which allows a remote microphone and preamplifier *to be utilized* without the loss of audio signal integrity. The input is AC-coupled and terminated by a 10-kilohm resistance. The useful audio bandwidth of the VRM is from 200 to 7000 Hz. Excellent *recognition* is attainable with the reduced telephone bandwidths.

Highly Accurate Real-Time Operation

The input speech is analyzed by a 16-channel spectrum analyzer and converted to a digital representation of the characteristics of the spoken input. *This digital data* is then *converted to a fixed-size pattern that preserves the* information content of the spoken inputs while discarding redundant features. During word training, these patterns are used to derive templates for each vocabulary item. These templates are used in the recognition process for comparison with incoming spoken words. Vocabulary templates are stored in an onboard random-access memory (RAM), while the processing algorithms are contained in an *onboard read-only memory* (ROM) operating in conjunction with a microprocessor.

The VRM has two *training modes*: (1) normal training in which the vocabulary storage is cleared and a new vocabulary is trained by speaking chosen words a selectable number of times, and (2) updating of word patterns in which the *stored reference* patterns for the specified vocabulary are augmented by additional training.

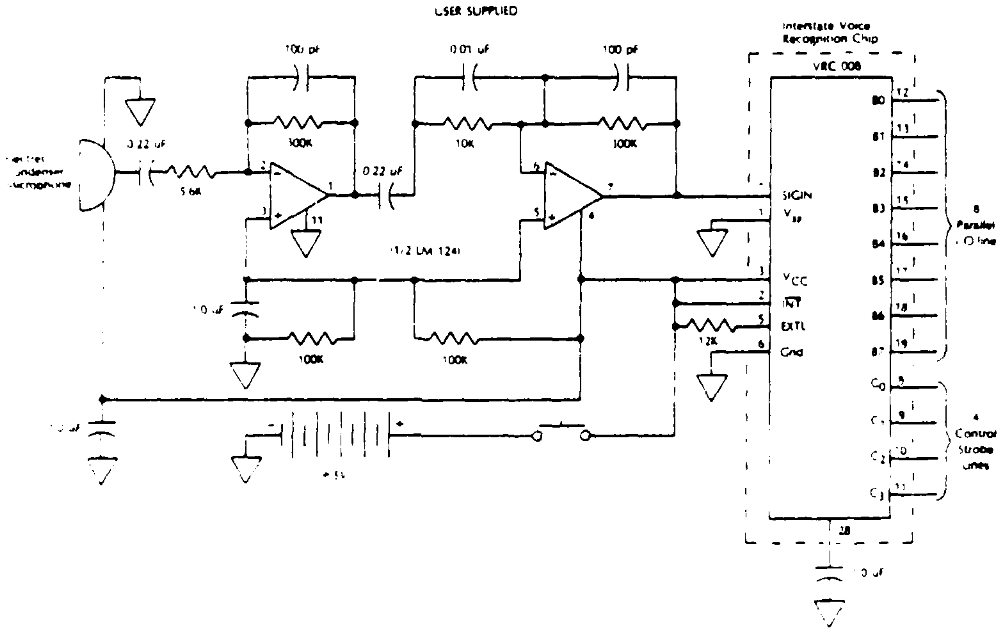
The VRM automatically rejects utterances during training that do not sufficiently agree with the same utterance from previous *training samples of that word. This prevents significant alteration of* a vocabulary reference pattern due to spurious noise (bumping the microphone, door closure, coughing, speaking inconsistencies, or simply failing to utter the vocabulary in the *specified sequence*). Thus, it may be necessary to repeat an utterance before being prompted to the next sequential utterance.

TABLE 9

Interstate Voice Recognition Chip VRC008

INTERSTATE ELECTRONICS CORPORATION

VOICE RECOGNITION CHIP Model VRC008 (advance information)



- **Low-cost voice recognition**
- **8-word vocabulary**
- **Speaker independent**
- **85% accuracy for general population**
- **Up to 95% accuracy for specific speakers**
- **NMOS and CMOS versions**
- **8 parallel I/O lines**
- **Single +5-volt power supply**

A Low-Cost 28-Pin, Single-Chip Voice Recognition System

Interstate's 28-pin single-chip VRC008 system employs a unique method for processing of analog speech data and recognition of spoken utterances.

Designed for a wide variety of high-volume consumer applications, this microcomputer provides low-cost voice control capability for appliances, toys, games, and other voice automation products. The system is speaker-independent and recognizes with high accuracy eight spoken words or phrases, translating verbal commands (i.e. "walk," "stop," "channel four," "turn right," etc.) into action via associated circuitry. In a typical application "wake up" activates the system into a receptive mode and prepares it to accept input speech; the word "relax" stops the system.

Programmable for a selected vocabulary, the VRC008 recognizes speech by detecting the state sequence of certain voiced and unvoiced parameters in the incoming word or phrase and comparing this sequence with the stored sequence of a prespecified vocabulary. With recognition accomplished, the system then outputs a bit pattern for the word number identified. The state sequence and recognition parameters are stored in the on-chip ROM.

Interstate customizes the VRC008 to specific user vocabularies. In this process the customer defines the particular functions to be performed by his product and IEC provides assistance in selecting a vocabulary suited to those functions.

Interstate markets the VRC100-1 voice recognition chip set. This chip set sells for $385. Basically, this set consists of two chips to be used as building blocks for speech recognition systems capable of recognizing as many as 100 words or short phrases. Tables 10 and 11 indicate a diagram of a suggested setup using the VRC100-1 chip set.

The VRC100-1 chip set nicely typifies Interstate's approach to speech recognition, which is briefly set forth below.

Speech input is analyzed by a 16-channel spectrum analyzer and converted to a digital representation of the characteristics of the spoken input. The digital data are then converted to fixed-size templates which preserve the information content of the spoken input while discarling redundant features. During training, stored patterns are used to derive templates for each word pattern. These templates are next used in the recognition process for comparison with incoming speech templates. Presumably, incoming templates are correlated with stored templates for actual word recognition. Vocabulary templates are stored in the external ROM, while processing algorithms are contained within the speech analyzer device.

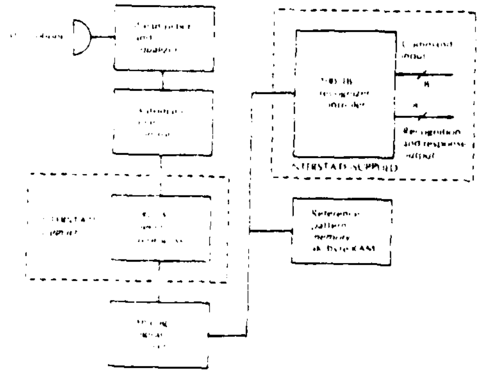
As this report was in preparation, Interstate added another voice recognition module to its voice input product line. This is the VRT300, which is reported to have a 100 word vocabulary. The unit is designed to be a single

TABLE 10

Interstate VRC100 1 Voice Recognition Chip Set

INTERSTATE ELECTRONICS CORPORATION

VOICE RECOGNITION CHIP SET Model VRC100-1 (advance information)


Command input
8
8
100-1B recognizer controller
Recognition and response output
INTERSTATE SUPPLIED
Reference pattern memory 4K Byte RAM

- **Speech recognition chip set**
- ***Highly accurate real-time operation (99% +)***
- **100-word vocabulary**
- **Trainable for any vocabulary in any language**
- **Two training modes — train and update (all or part of the vocabulary)**
- **Usable with direct microphone, wireless communication, or telephone**
- **Selectable decision threshold for rejection of unwanted inputs**
- **Input/output port configuration for easy product integration**

High-Accuracy Voice Recognition in a Chip Set

Interstate's Model VRC100-1 voice recognition chip set consists of two chips used as building blocks for speech recognition systems capable of recognizing as many as 100 words or short phrases (see illustration). These two integrated circuits are designated 100-1A and 100-1B.

The 100-1A chip is a 28-pin integrated circuit providing audio-spectrum analysis over the range of intelligibility for speech (200 to 7000 Hz). The analog input to the 100-1A is 5 volts rms maximum from a low-output impedance source. The 100-1A consists of 16 bandpass filters, each followed by a half-wave rectifier and a second-order low-pass filter with 25-Hz cutoff. The monolithic 100-1A utilizes NMOS switched-capacitor technology with 80 operational amplifiers to achieve the required audio-spectrum analysis. Additionally, this chip contains a 16-channel analog multiplexer and decoder which require timing signals from a single TTL 1-MHz clock. The analog multiplexer is addressed via four TTL lines. The analog output of the 100-1A chip is from a buffer amplifier. This output is suitable for a 0- to 5-volt user-supplied analog-to-digital converter.

The 100-1B chip is Interstate's 40-pin recognizer controller. This chip contains the entire algorithm for recognition of isolated speech utterances including (1) word boundary detection, (2) amplitude normalization, (3) end point time compression, and (4) programmable vocabulary syntax. The 100-1B provides parallel I/O and control of the analog multiplexer and the analog-to-digital converter. Commands provided via the parallel input port are interpreted by the 100-1B chip. Recognition and command responses are provided via the parallel output port. All data I/O is in the form of ASCII characters.

Efficient, Real-Time Performance

Speech input is analyzed by a 16-channel spectrum analyzer and converted to a digital representation of the characteristics of the spoken input. This digital data is then converted to a fixed-size pattern that preserves the information content of the spoken inputs while discarding redundant features. During word training, these patterns are used to derive templates for each vocabulary item. The templates are then used in the recognition process for comparison with incoming spoken words. Vocabulary templates are stored in the external RAM, while the processing algorithms are contained within the speech analyzer device.

The ROM accommodates eleven user commands. These include two training modes: (1) normal training in which all or part of the specified vocabulary is cleared and then trained a selectable number of samples, and (2) updating of word patterns in which the stored reference patterns of the specified vocabulary are augmented by additional training.

The VRC100-1 training algorithm automatically rejects utterances during training that do not sufficiently agree with the same utterance from previous training samples of the word. This prevents significant alteration of a vocabulary reference pattern caused by spurious noise (bumping the microphone, door closure, coughing), speaking inconsistencies, or simply failing to utter the prompted vocabulary item. In such an event, it may be necessary to repeat an utterance before being prompted to the next sequential utterance.

TABLE 11

Interstate VRC100-1 Chip Set Further Details

Additional commands allow the VRC100-1 to upload or download reference patterns via the selected I/O port. The reset command is used to initialize the VRC100-1 chip set RAM and to define I/O mode and format. The VRC100-1 chip set also allows control of the rejection of invalid utterances via the set reject and read reject threshold commands.

An additional command allows programmable control of the analog-to-digital reference voltage and preamplifier gain. Finally, the major operational command of the VRC100-1 is the recognize command. This command allows recognition of any specified vocabulary up to 100 words. The command allows recognition of both contiguous and/or random syntax with one or more common subvocabularies.

A Family of Voice Recognition Products

The Model VRC100-1 chip set and other Interstate-developed chip sets benefit both OEMs and end users by enabling the design flexibility to support a wide range of applications. Interstate's family of speech recognition products can be economically incorporated into a variety of industrial systems and consumer products – from large-scale inventory control equipment to personal computers and hobby items.

MODEL VRC100-1 SPECIFICATIONS

Performance

Vocabulary Size: Up to 100 isolated words and/or phrases.
Percent Recognition Accuracy: 99+ percent.
Reject Threshold: User-selectable.
Longest Utterance Duration: 1.25 second.
Minimum Between-Word Pauses (User Selectable): 160 milliseconds
Minimum Word Length (User Selectable): 80 milliseconds.
Approximate Response Time: (50 + 2N) milliseconds, where N = active vocabulary size with a 4 MHz crystal.

Host Commands

1. Train
2. Update
3. Reset
4. Set reject threshold
5. Read reject threshold
6. Download
7. Upload
8. Set analog-to-digital preamplifier gain
9. Recognize
10. Write parameters
11. Read parameters

Input/Output

Parallel TTL input/out, eight data input bits, eight data output bits with four control lines. All data input/output is in the form of ASCII characters.

Mechanical

Speech Preprocessor Chip: Dual in-line 28-pin package.
Recognizer/Controller Chip: Dual in-line 40-pin package.

Electrical

Power Requirements: 100-1A: +10V, −10V at 30 mA; 100-1B: +5 Vdc at 240 mA.

Interstate Electronics Corporation reserves the right to make changes to any products herein to improve reliability, function or design. Interstate does not assume any liability arising out of the application or use of any product or circuit described herein; neither does it convey any license under its patent rights nor the rights of others.

INTERSTATE
ELECTRONICS
CORPORATION

Voice Products Operations

1001 E. Ball Road, P.O. Box 3117, Anaheim, California 92803
Telephone 714/635-7210 TWX 910-591-1197 Telex 655443 & 655419
Call toll-free: in the continental U.S. 800/854-6979; in California 800/422-4580

plug-in board for use in the DEC VT100 terminals and similar models. It basically transforms spoken words into ASCII strings and transmits these strings to the host computer. The module costs $1,295. It was announced in the July 19,1982, issue of Computerworld.

Interstate has recently expanded their product lines in the areas of voice input and output devices substantially. Their whole approach typifies the increasing emphasis manufacturers are now placing on speech products in general. The consumer is finding new devices on the market faster than ever before, and this trend is expected to continue to accelerate.

Nippon Electric Company

Nippon Electric's speech recognition products may be classified as being in the relatively higher-priced, systems-level product category.

Nippon Electric's DP-100 speech recognizer gained wide attention for its ability to recognize connected speech. Their newest recognizer, the DP-200, is generally similar to the DP-100 but at a reduced price.

The DP-200 uses dynamic programming to match templates obtained during training with those from incoming speech data. One very noteworthy point about the DP-200 is, again, its ability to recognize connected speech. It also has a larger vocabulary than the DP-100 (150 words vs. 120 words).

The DP-200 is approximately 1/3 the size of the DP-100. The price of the DP-200 is approximately 20-30% less than that for the DP-100, which would make it approximately $33,000.

There are several further comments to make regarding the DP-200:

1) The DP-200 will recognize dialects.

2) Minimal training is required; one pass for most words, two for numerics.

3) The DP-200 uses dynamic programming to "warp" time frames of incoming speech to achieve best matching of words in the shortest time possible.

4) Optional audio response is available.

5) The DP-200 has a wider range of interfacing capabilities than the DP-100.

A final point to mention regarding the DP-200 is that it is a stand-alone system of a fairly compact nature, consisting of a speech recognition terminal, a remote control terminal, and a noise-cancelling microphone. Following is Table 12 which contains a brief description of the DP-200's approach to speech recognition, in addition to a block diagram of the Nippon system.

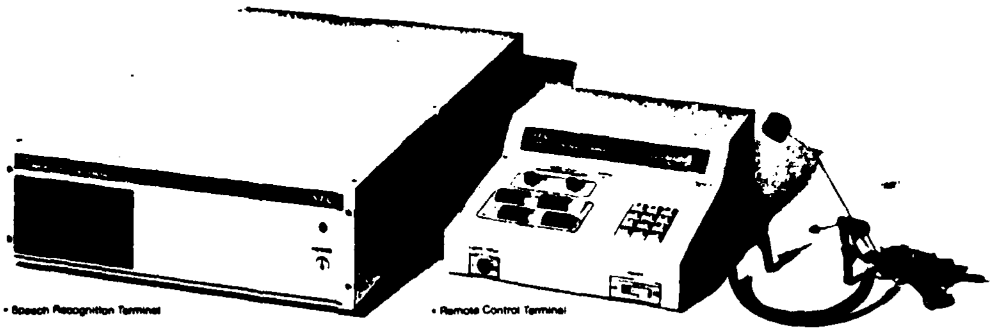
Scott Instruments

Scott Instruments markets terminal or systems level speech recognition products. Scott Instruments' voice entry terminal does not come with a host computer; this points out

TABLE 12

Nippon Electric DP-200

CSR:
Compatible Human-to-Machine Communications

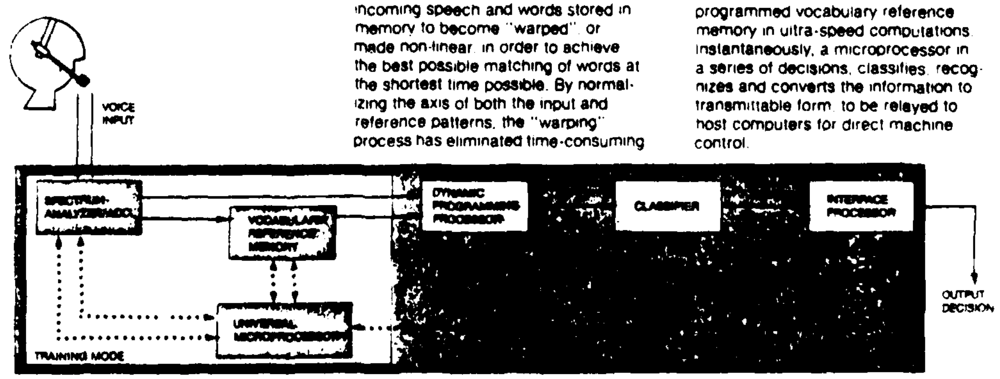


With the NEC DP-200 CSR technology has at last created the ideal way of dealing with machines, using and controlling them. The results are *striking—data entry is as simple as* sitting down and speaking in a normal conversational tone, while a computer captures your words instantly. What was once deemed technically impossible is now a reality. With the DP—200 CSR, NEC's research team has not only produced a quality data entry system but also created the most direct and efficacious way yet for man to use and control machines.

Until now, machines could neither recognize nor process speech patterns *that varied both in speed and in context.* Previous linear computers were stumped, unable to recognize where one word ended and the next one began. The thousands of variations contained in normal connected speech were deemed too great a task to be logically identified and stored by a machine.

The NEC CSR uses Dynamic Programming (DP), a time normalization method which has effectively solved these once unsurmountable problems. DP allows incoming speech and words stored in memory to become "warped", or made non-linear, in order to achieve the best possible matching of words at the shortest time possible. By normalizing the axis of both the input and reference patterns, the "warping" process has eliminated time-consuming errors due to word segmentation and incorrect matching. DP has provided the technological leap forward necessary to achieve direct and *compatible human-to-machine* communications.

The heart of the DP-200 lies in a series of high-speed computations utilizing Dynamic Programming techniques. Incoming speech signals from the operator's microphone (as analog waveforms) pass through a spectrum analyzer and are immediately converted to a digital signal. The signal is transferred and compared to the pre-programmed vocabulary reference memory in ultra-speed computations. Instantaneously, a microprocessor in a series of decisions, classifies, recognizes and converts the information to transmittable form, to be relayed to host computers for direct machine control.



the fact that drawing the line between systems level products and board level products is not always a clear cut distinction.

Scott Instruments markets one basic recognizer, the VET-2 Voice Entry Terminal. The basic system consists of the VET-2 preprocessor, software and demonstration programs, operations manual, and a noise-cancelling microphone.

The VET-2 is available for Apple or TRS-80 computers. It is noted that the unit interfaces with off-the-shelf software, or programs may be written in BASIC, INTEGER-BASIC, APPLESOFT-BASIC, or machine code.

The VET-2 has a 40-word basic vocabulary, with an overlay feature to allow access to additional vocabulary residing in disk storage. The Vet-2 claims an accuracy rate of 98%+. A five or six training pass approach is suggested.

Scott Instruments' approach utilizes an acoustic preprocessor to analyze the acoustic signal within a range of 300-4000 Hz. Analysis consists of breaking the frequency range down into 2 regions (300-1000 Hz. and 1000-4000 Hz.), then taking zero-crossing measures in both regions and extracting the amplitude envelopes of the two regions. The four resulting analog data lines are converted to digital form at the request of the host computer.

Words for the VET-2 can be up to 1.5 sec. in duration, and up to 20 characters long. The template area for a 40-word vocabulary requires approximately 4600 bytes of storage. Control software requires approximately 6000 bytes for a total of 10.6K memory required in the host computer. Following is Table 13 with operation specifications and key features of the VET-2 which retails for a relatively inexpensive $795. One the most difficult areas to show comparison of recognizers is in price, for the capabilities vary as a function of cost. Nontheless, cost is a major consideration to most buyers.

Threshold Technology

Threshold Technology's speech recognition products are generally geared toward the high end of the market; their speech recognizers are characterized as systems level products.

Threshold Technology has just started a new subsidiary, called Auricle. Together, these two groups market three basic lines of speech recognizers. The two Threshold recognizers, the 580 and 680 units, are among a select few recognizers that will accept connected speech input. Threshold's approach uses dynamic programming, where words, rather than combinations of phonemes, are recognized as units. Threshold's Votrax unit uses a 16-channel bandpass filter with a rectified compressor with proprietary circuitry and a commercial codec.

Threshold emphasizes that its units will accept connected words with very short interword pauses. Threshold calls this feature "Quiktalk". In fact, Threshold notes

TABLE 13

Scott Instruments

KEY FEATURES

Available on the APPLE or TRS-80

Easy to use---interfaces with off-the-shelf software or programs may be written in BASIC, INTEGER-BASIC, APPLESOFT-BASIC or MACHINE CODE

KEYVET feature allows voice to be used in conjunction with the keyboard (Apple only)

Multiple user capabilities with no increase in storage requirements

40-word vocabulary with overlay feature to access additional vocabularies from disk

High accuracy (98%+)

SPECIFICATIONS

Requires an Apple II or Apple II-plus, 48K machine with at least one disk drive, or a TRS-80 Model I with 32K or 48K and two disk drives

SIZE: Approximately 1¼" H x 8" W x 11" D

WEIGHT: Approximately 5 lbs

POWER: Apple power supply or TRS-80---115 VAC. 60 Hz. 15 Watts that it will accept words faster than normally encountered in continuous speech (180 wpm). It is noted that this feature permits data entry much faster than via keyboard entry, with a claimed accuracy of 99%+.

Prices for units with the Quiktalk feature should be, Threshold states, 10-20 % above those for the older 500, 600, and VIP-100 units, which can be upgraded to include this feature. Thus, the 580 unit sells for approximately $16,000; the 680 unit sells for approximately $13,300.

The Threshold 580 recognizer has a 60-word or phrase vocabulary (expandable to 340). It includes two noise-cancelling microphones. It produces ASCII coded output and has a 16-character alphanumeric display for voice data entry and verification. Also included are ready and reject indicators. The unit has a reject decision level (externally set by program control). The 580 accepts words/phrases up to two seconds in duration.

The Threshold 680 recognizer features a 50-word or phrase vocabulary, and also produces ASCII coded output. It permits local tape cartridge storage of user speech patterns, training prompts, and output messages. The 680 includes a CRT display terminal for operator prompting, editing, and verification. The unit is current loop output compatible from 50 Baud to 19.2 Baud. The 680 has optional wireless radio input. It accepts words or phrases up to two seconds in duration, like the 580. Finally, for a quick comparison of features of the 580/680 recognizers, Table 14 gives a short listing of comparative specifications.

Threshold 580/680 units appear to utilize proprietary filter algorithms which statistically match digitized input speech data with stored templates. Threshold states that its Quiktalk feature consists of recognizing strings of words as units, rather than recognizing individual words. This permits detection of the shortest possible pauses between words. Thus both of these units nearly attain the long-sought goal of continuous speech recognition.

Further specifications are given in Tables 15 and 16 for the Threshold 580 and 680 recognizers.

The Auricle-I is designed to be a lower cost recognizer, with a vocabulary of 80 words or short phrases. It includes LSI circuitry, boasts an accuracy rate over 99%, and includes a settable reject level. One purpose of the Auricle-I is to function as a benchtop development system which will help familiarize designers with speech recognition and help them decide if such an approach is suitable for their end products. The Auricle-I costs $2,500. Also under development is a board-level product, the Auricle-II, which is a speech recognizer card.

The Auricle-I uses a 16-channel bandpass filter and a rectifier/compressor that consists of proprietary circuitry and a commercial codec. The Auricle's host Z-80 correlates input templates with stored templates to determine matches.

TABLE 14

Threshold Technology 580/680 Recognizers

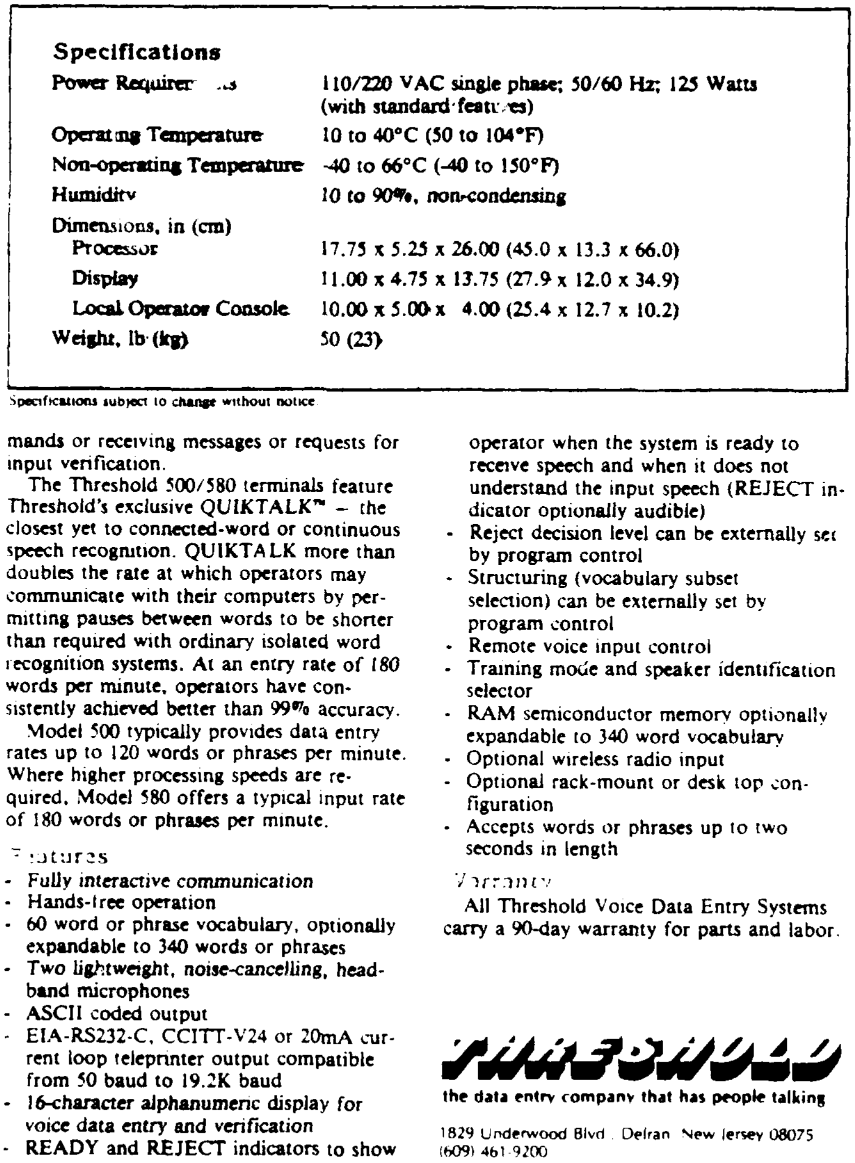
-- COMPARATIVE FEATURES OF BOTH SYSTEMS ARE SHOWN BELOW --

	THRESHOLD 680	THRESHOLD 580
Operating speeds (@ 99% accuracy)	180 words/minute	180 words/minute
Vocabulary	40 words expandable to 250 words	60 words expandable to 370 words
Display	CRT	16-character alphanumeric
Speaker Training Data	Stored in local tape cassette	Stored in host computer
Output Code for Each Utterance	User programmable character or string of characters; stored on local tape cassette	Unique ASCII code
Vocabulary Display Message for Operator Prompting	User programmable; stored on local tape cassette	Controlled by host computer
Electrical Interface	Standard RS232C or current loop; serial asynchronous ASCII	Standard RS232C or current loop, serial asynchronous ASCII
Software Compatibility with Standard Teletype Terminal	Fully compatible and no special software required	Requires special host computer software to handle communications protocol

(Rev. 3-14-80)

TABLE [illegible]

Threshold 580 Specifications



Specifications

Power Requirements	**110/220 VAC single phase; 50/60 Hz; 125 Watts (with standard features)**
Operating Temperature	10 to 40°C (50 to 104°F)
Non-operating Temperature	-40 to 66°C (-40 to 150°F)
Humidity	10 to 90%, non-condensing
Dimensions, in (cm)	
Processor	17.75 x 5.25 x 26.00 (45.0 x 13.3 x 66.0)
Display	11.00 x 4.75 x 13.75 (27.9 x 12.0 x 34.9)
Local Operator Console	10.00 x 5.00 x 4.00 (25.4 x 12.7 x 10.2)
Weight, lb (kg)	50 (23)

Specifications subject to change without notice

mands or receiving messages or requests for input verification.

The Threshold 500/580 terminals feature Threshold's exclusive QUIKTALK™ – the closest yet to connected-word or continuous speech recognition. QUIKTALK more than doubles the rate at which operators may communicate with their computers by permitting pauses between words to be shorter than required with ordinary isolated word recognition systems. At an entry rate of 180 words per minute, operators have consistently achieved better than 99% accuracy.

Model 500 typically provides data entry rates up to 120 words or phrases per minute. Where higher processing speeds are required, Model 580 offers a typical input rate of 180 words or phrases per minute.

Features

- Fully interactive communication
- Hands-free operation
- 60 word or phrase vocabulary, optionally expandable to 340 words or phrases
- Two lightweight, noise-cancelling, headband microphones
- ASCII coded output
- EIA-RS232-C, CCITT-V24 or 20mA current loop teleprinter output compatible from 50 baud to 19.2K baud
- 16-character alphanumeric display for voice data entry and verification
- READY and REJECT indicators to show operator when the system is ready to receive speech and when it does not understand the input speech (REJECT indicator optionally audible)
- Reject decision level can be externally set by program control
- Structuring (vocabulary subset selection) can be externally set by program control
- Remote voice input control
- Training mode and speaker identification selector
- RAM semiconductor memory optionally expandable to 340 word vocabulary
- Optional wireless radio input
- Optional rack-mount or desk top configuration
- Accepts words or phrases up to two seconds in length

Warranty

All Threshold Voice Data Entry Systems carry a 90-day warranty for parts and labor.

THRESHOLD

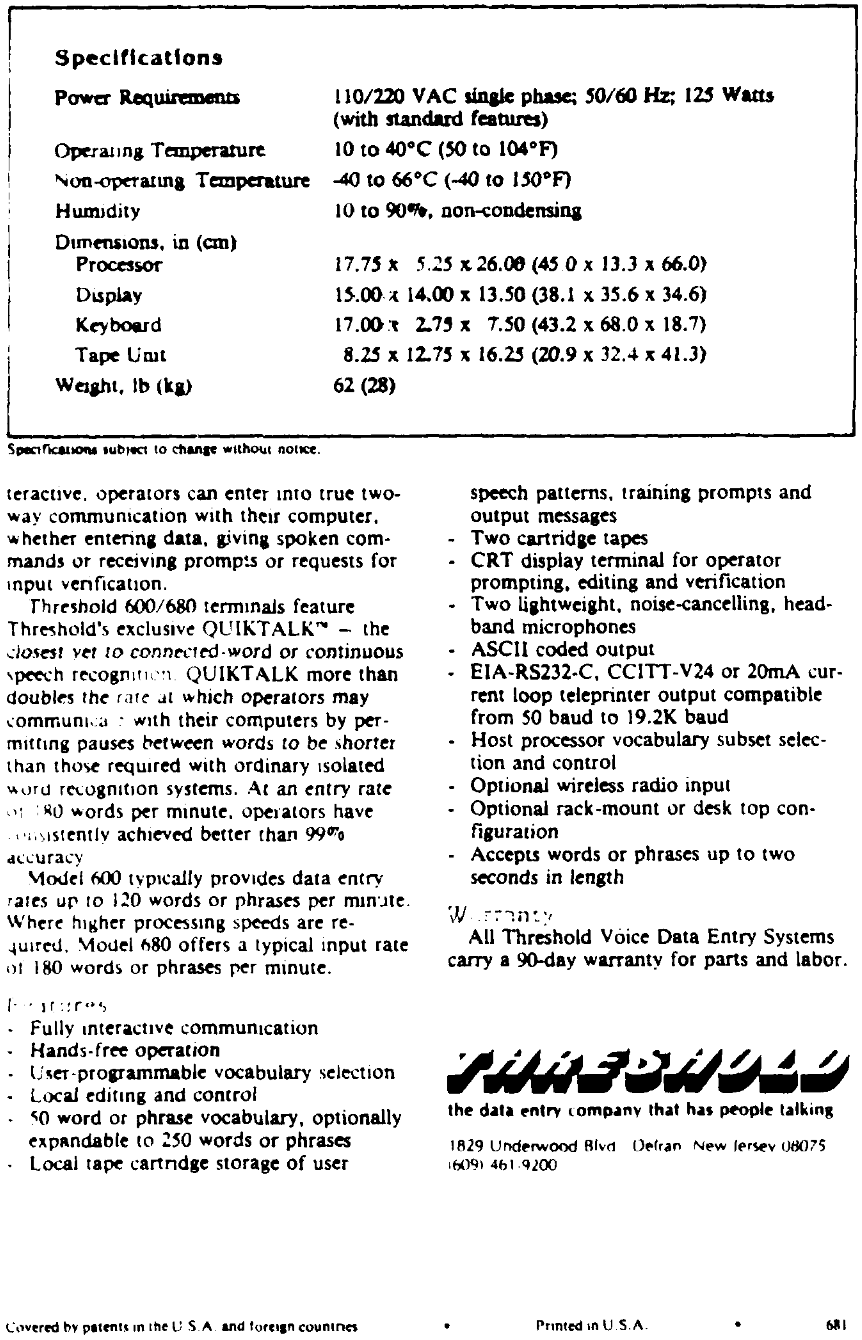
the data entry company that has people talking

1829 Underwood Blvd., Delran, New Jersey 08075
(609) 461-9200

Covered by patents in the U.S.A. and foreign countries • Printed in U.S.A. • 681

TABLE 16

Threshold 680 Specifications



Specifications

Power Requirements	**110/220 VAC single phase; 50/60 Hz; 125 Watts (with standard features)**
Operating Temperature	10 to 40°C (50 to 104°F)
Non-operating Temperature	-40 to 66°C (-40 to 150°F)
Humidity	10 to 90%, non-condensing
Dimensions, in (cm)	
Processor	17.75 x 5.25 x 26.00 (45.0 x 13.3 x 66.0)
Display	15.00 x 14.00 x 13.50 (38.1 x 35.6 x 34.6)
Keyboard	17.00 x 2.75 x 7.50 (43.2 x 68.0 x 18.7)
Tape Unit	8.25 x 12.75 x 16.25 (20.9 x 32.4 x 41.3)
Weight, lb (kg)	62 (28)

Specifications subject to change without notice.

teractive, operators can enter into true two-way communication with their computer, whether entering data, giving spoken commands or receiving prompts or requests for input verification.

Threshold 600/680 terminals feature Threshold's exclusive QUIKTALK™ – the *closest yet to connected-word or continuous speech recognition.* QUIKTALK more than doubles the rate at which operators may communicate with their computers by permitting pauses between words to be shorter than those required with ordinary isolated word recognition systems. At an entry rate of 180 words per minute, operators have consistently achieved better than 99% accuracy

Model 600 typically provides data entry rates up to 120 words or phrases per minute. Where higher processing speeds are required, Model 680 offers a typical input rate of 180 words or phrases per minute.

Features
- Fully interactive communication
- Hands-free operation
- User-programmable vocabulary selection
- Local editing and control
- 50 word or phrase vocabulary, optionally expandable to 250 words or phrases
- Local tape cartridge storage of user speech patterns, training prompts and output messages
- Two cartridge tapes
- CRT display terminal for operator prompting, editing and verification
- Two lightweight, noise-cancelling, headband microphones
- ASCII coded output
- EIA-RS232-C, CCITT-V24 or 20mA current loop teleprinter output compatible from 50 baud to 19.2K baud
- Host processor vocabulary subset selection and control
- Optional wireless radio input
- Optional rack-mount or desk top configuration
- Accepts words or phrases up to two seconds in length

Warranty

All Threshold Voice Data Entry Systems carry a 90-day warranty for parts and labor.

THRESHOLD

the data entry company that has people talking

1829 Underwood Blvd Delran New Jersey 08075
(609) 461-9200

Covered by patents in the U.S.A. and foreign countries • Printed in U.S.A. • 681

The Auricle-I has a 40-word vocabulary; this vocabulary can be enlarged by using a host computer's memory to store templates.

The Auricle-I requires three-pass training for vocabulary items. The unit limits vocabulary to a very low number of highly differentiated responses, so it is possible to program in templates with a wide variety of pronunciations. Response time for the Auricle-I is listed at 350 msec. for a word of less than 1.2 seconds in duration. It is noted that the Auricle-I is a completely stand-alone system with its own power supply and noise-cancelling microphone. The unit sells for $2,480. Further information on Auricle-I is given in Table 17.

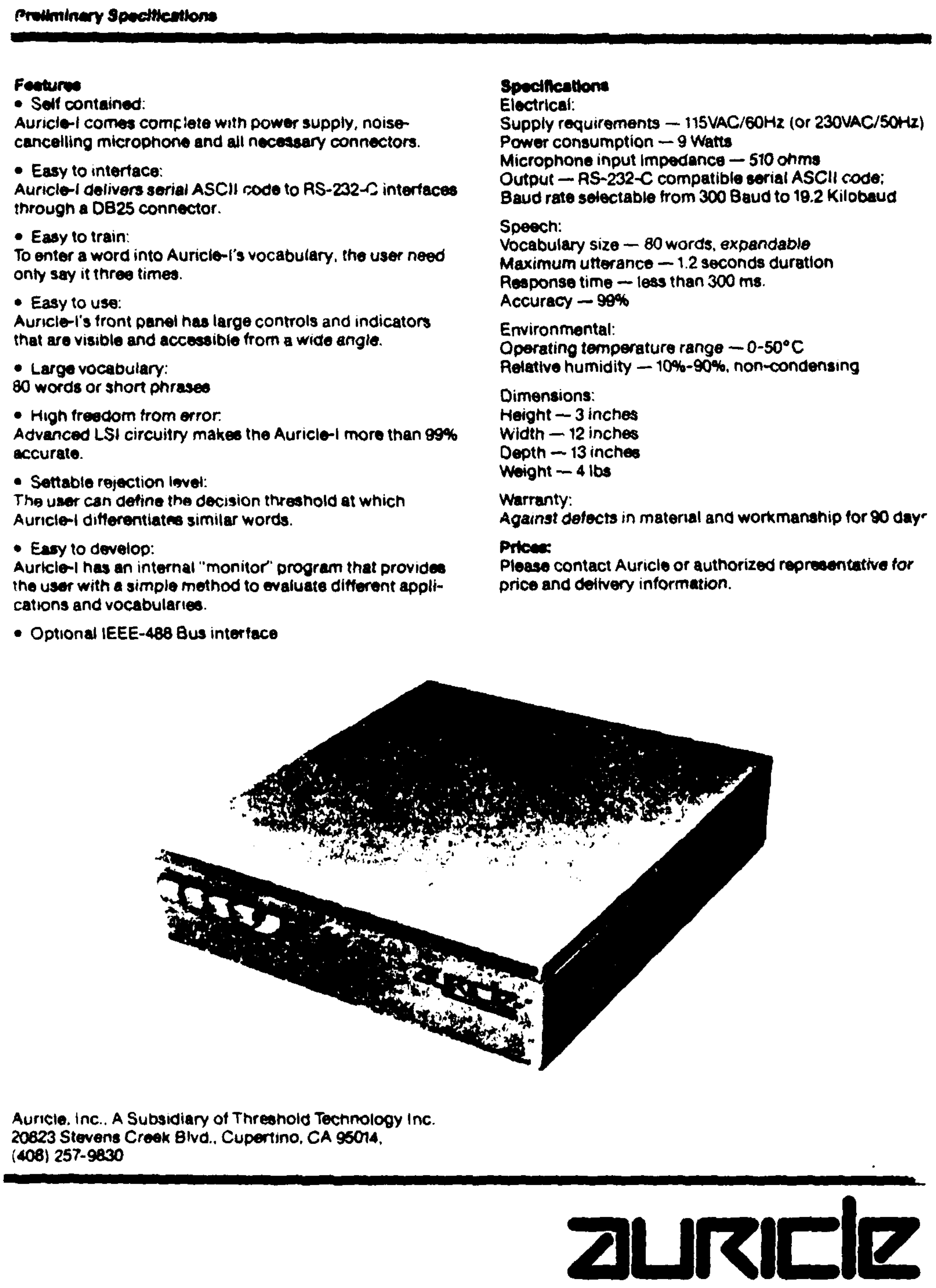
Verbex

Verbex speech recognition products fall into the high end systems level category.

Verbex is one of the oldest manufacturers of speech recognizers, and has been a pioneer in the area of speaker-independent speech recognition technology. Currently, Verbex markets two speech recognizers. The Verbex Model 1800 is an isolated word, speaker-independent speech recognizer (multi-user). The 1800 comes with a recognition vocabulary consisting of the 10 digits, "zero" through "nine", plus "yes/no". This vocabulary can be expanded to 50 words. The 1800 includes voice response (32

TABLE 17

Auricle-I Recognition Systems



Preliminary Specifications

Features

- Self contained:
Auricle-I comes complete with power supply, noise-cancelling microphone and all necessary connectors.
- Easy to interface:
Auricle-I delivers serial ASCII code to RS-232-C interfaces through a DB25 connector.
- Easy to train:
To enter a word into Auricle-I's vocabulary, the user need only say it three times.
- Easy to use:
Auricle-I's front panel has large controls and indicators that are visible and accessible from a wide angle.
- Large vocabulary:
80 words or short phrases
- High freedom from error:
Advanced LSI circuitry makes the Auricle-I more than 99% accurate.
- Settable rejection level:
The user can define the decision threshold at which Auricle-I differentiates similar words.
- Easy to develop:
Auricle-I has an internal "monitor" program that provides the user with a simple method to evaluate different applications and vocabularies.
- Optional IEEE-488 Bus interface

Specifications

Electrical:
Supply requirements — 115VAC/60Hz (or 230VAC/50Hz)
Power consumption — 9 Watts
Microphone input impedance — 510 ohms
Output — RS-232-C compatible serial ASCII code;
Baud rate selectable from 300 Baud to 19.2 Kilobaud

Speech:
Vocabulary size — 80 words, *expandable*
Maximum utterance — 1.2 seconds duration
Response time — less than 300 ms.
Accuracy — 99%

Environmental:
Operating temperature range — 0-50°C
Relative humidity — 10%-90%, non-condensing

Dimensions:
Height — 3 inches
Width — 12 inches
Depth — 13 inches
Weight — 4 lbs

Warranty:
Against defects in material and workmanship for 90 day

Prices:
Please contact Auricle or authorized representative for price and delivery information.

Auricle, Inc., A Subsidiary of Threshold Technology Inc.
20823 Stevens Creek Blvd., Cupertino, CA 95014,
(408) 257-9830

auricle

words or 16 seconds of speech); this can be expanded up to 512 words or 256 seconds of audio response. Both the recognition and response vocabularies can be customized as required for individual applications. The 1800 can also be used over the telephone. Table 18 indicates the Verbex Model 1800's basic specifications.

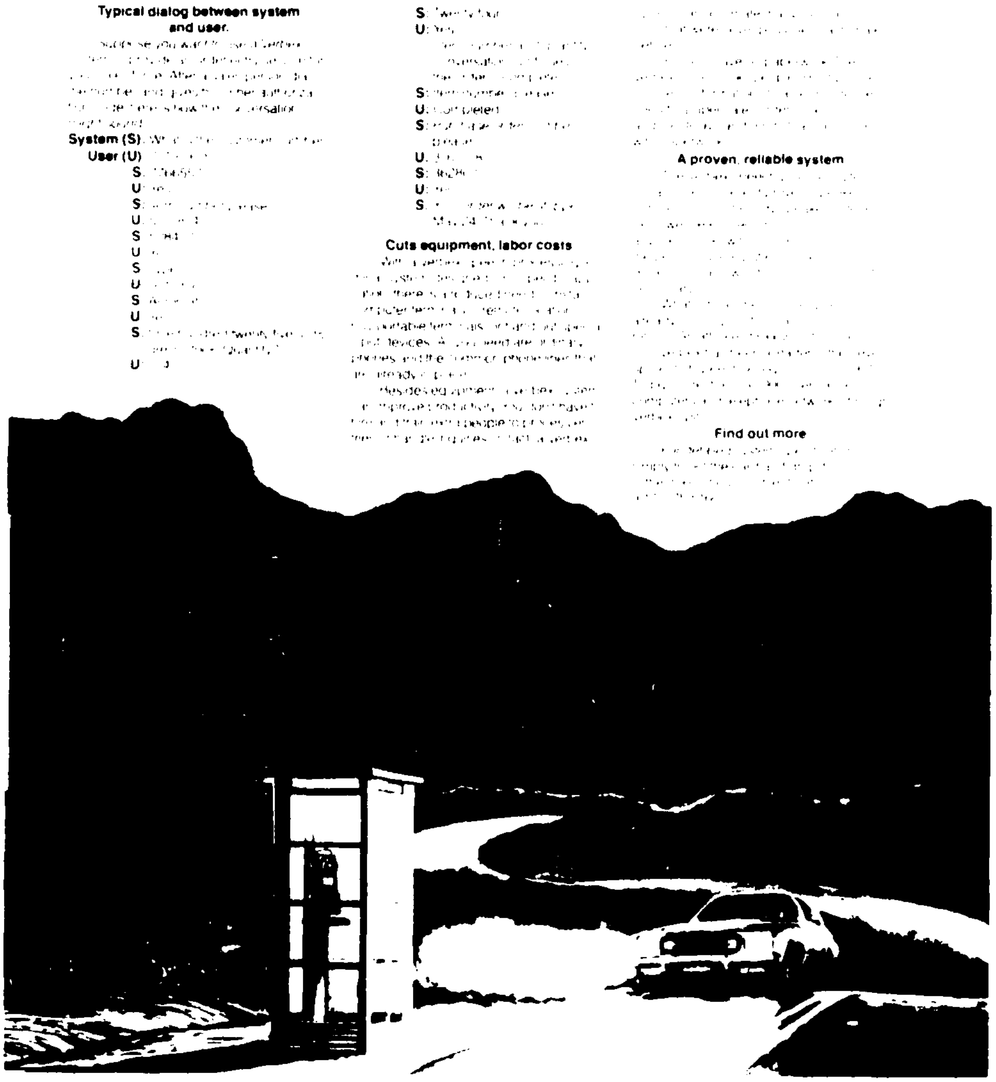
The second speech recognizer marketed by Verbex is the Verbex Model 1800-CSRS. This unit is speaker-dependent and handles continuous speech input. The unit also has single-channel entry microphone input; it is basically designed for high accuracy digit entry, plus 10 isolated command words. We wish to point out the further possibility of using the Verbex Model 1800 recognizer to spot keywords in incoming Coast Guard radio transmissions. This unit is designed to accept isolated words as input, but this may be sufficient for the Coast Guard's needs. That is, we suspect that keywords, such as "mayday," may actually be pronounced slowly enough for the Verbex unit to recognize these with high accuracy. One unknown in this area concerns how the Verbex unit might generate "false alarms" for connected speech input, from which relatively "isolated" keywords would have to be separated by the recognition unit.

Voicetek

Voicetek's product line of speech recognition products may be categorized as board level. With their relatively

TABLE 18

Verbex Model 1800 Used With The Telephone



Typical dialog between system and user.

[illegible]

System (S): [illegible]

User (U): [illegible]

S: [illegible]

U: [illegible]

S: [illegible]

U: [illegible]

S: [illegible]

U: [illegible]

S: [illegible]

U: [illegible]

S: [illegible]

U: [illegible]

S: [illegible]

U: [illegible]

S: Twenty four

U: Yes

[illegible]

S: [illegible]

U: Completed

S: [illegible]

U: [illegible]

S: [illegible]

U: [illegible]

S: [illegible]

Cuts equipment, labor costs

[illegible]

A proven, reliable system

[illegible]

Find out more

[illegible]

inexpensive board level recognizers, Voicetek is seeking the home computer, hobbyist market.

Voicetek markets a line of inexpensive voice input/output devices for small computer systems. Voicetek notes that they have been able to achieve inexpensive prices for their voice I/O devices due to their having successfully compressed required electronics onto a single integrated circuit chip. Following is a list of major features regarding Voicetek's voice I/O devices, which are called Cognivox units:

1) Unlike speech recognizers that employ frequency domain (filter bank) analysis, Cognivox units operate on the time-domain signal. This allows for high performance at low cost. Cognivox units also use a new and exclusive nonlinear pattern matching algorithm to enhance performance. Voicetek technology does involve the use of Fast Fourier Transforms (FFT), but details are not available in this area.

2) Voicetek units have been given a 50 hour burn-in or testing period.

3) A Cognivox unit is priced lower than either a comparable speech recognizer or a voice-response unit, yet it combines both features.

4) A Cognivox unit features easy training, with the user repeating the desired vocabulary three times at the prompting of the host computer.

Voicetek's speech products are basically divided into three lines:

1) VIO-1000 series of voice I/O peripherals. These are priced at $249 and are for Rockwell AIM-65, PET/CBM 16K or 32K, or Apple II computers. This is the top-of-the-line Voicetek unit.

2) VIO-XXX series of voice I/O peripherals. These are priced at $149 and are for economical voice I/O applications that do not require high fidelity speech output. The VIO-XXX is suitable for Exidy's Sorcerer, Z-80 based systems, TRS-80, LII, 16K, and PET/ CBM, 16K or 32K computers.

3) SR-100A and SR-100P units. These are speech recognition peripherals for the AIM-65 (4K) as well as the PET/CBM (8K, 16K, and 32K) computers.

Finally, note that Voicetek software is written in BASIC on cassette. It has filtering routines, including FFT. The following Table 19 summarizes major Voicetek features.

Votan

Votan manufactures both systems and board level recognition products.

Votan has a very interesting approach to voice I/O, which consists of a reversible algorithm that works for both voice input and voice output.[1] Their top-of-the-line model

[1] Voice output is an imminent enhancement.

TABLE 19

VOICETEK COGNIVOX

1. Speech input and output combined in one unit.
COGNIVOX is the only unit in the market that allows both speech input and output. Our experience with COGNIVOX and customer feedback indicates that this is the way to go in speech peripherals.

2. Exter.sive applications software and support.
This is the area of crucial importance since, for most users, the utility of a given system is directly proportional to the available applications software. Recognizing this reality, VOICETEK, has a strong commitment to seek out and develop applications for speech I/O. We currently offer two sophisticated speech-operated video games, as well as utilities such as a talking calculator, vocal memory dump, etc. We are also working on a series of application articles to be published in the major microcomputing magazines, such as BYTE, Kilobaud and Creative Computing. Applications include speech-operated instruments, voice-controlled machines and toys, talking appliances that respond to spoken commands, and so on.

3. State-of-the-art design.
Unlike other speech recognizers that employ frequency domain analysis, COGNIVOX, operates on the time-domain signal. This novel and unique approach allows for high performance at low cost. In addition, COGNIVOX employs a new and exclusive non-linear pattern matching alogrithm that significantly enhances its performance.

4. Quality hardware.
The COGNIVOX hardware is carefully designed and assembled. It is tested after assembly and again after a 50-hour bum-in period to insure long and trouble-free life. The COGNIVOX hardware is enclosed in a beautiful injection-molded instrument case, giving it an elegant appea appearance.

5. Affordable price.
COGNIVOX is priced lower than either a comparable speech recognizer or a voice-response unit, yet it combines both features. The low price is made possible by innovative design and by our conviction that voice I/O must be priced right before it gains the wide appeal it deserves.

6. Easy Training.
Today's technology allows only speaker-dependent recognizers, meaning that the recognizer must be trained to the voice of the individual user. In the case of COGNIVOX, this training is very easy and can be done quickly, as the user must repeat the vocabulary three times at the prompting of the computer. Training the voice response portion is also very simple, requiring that the user pronounce the voice response vocabulary only once.

is the Votan V1000. The V1000 is a stand-alone development system designed to enable product planners to evaluate the use of Votan's speech technology in proposed or existing products or systems. This is an increasingly popular approach with voice I/O device manufacturers in general. The V1000 sells for $5,000. Votan also markets the V2000, which is an industrial control module, with training and display functions carried out by the host software. It has a list price of $4,400. Finally, Votan markets the V3000 which is an O.E.M. circuit board, at a cost of $3,000.

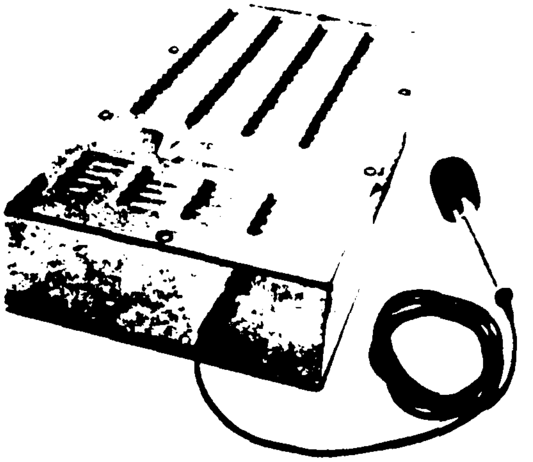
The V1000 is a stand-alone device. It will accept words or phrases up to two seconds maximum duration. It has a capacity of up to 100 seconds word storage (approximately 160 words single-trained or 80 words double-trained). The V1000 will operate under very high noise conditions (up to 85dB of background noise).

Votan uses an analog-to-digital converter to transform incoming speech data into a digital representation. A proprietary algorithm then processes the speech signal into its frequency components. Following the spectral transformation, dynamic programming warps spectral templates for comparison with reference templates. The spectral processing algorithm is reversable. This means that it should be able to accommodate speech synthesis as well as speech recognition.

Votan's synthesis chip provides user programmability, which is lacking in other LPC-based synthesis chips. Thus Votan's chip should be user-trainable in the field for easy accommodation of new vocabulary for synthesis. This approach should also allow new flexibility in speech selection for synthesis; previous LPC synthesizers have had to be reprogrammed in the laboratory. Votan promises significant future enhancements to their system. Following is Table 20, listing these enhancements, plus the general operating specifications of the V1000.

TABLE 20

Votan V1000



FUTURE ENHANCEMENTS

In a continuing effort to bring you state of the art recognition and synthesis capabilities, Votan is currently augmenting its technology to incorporate

Speaker Independent Recognition

This feature will allow our products to recognize a vocabulary word spoken by any user without individual user training

Speech Synthesis

Votan's integrated technology uses the same proprietary algorithms for both synthesis and recognition

Continuous Speech Input

Special algorithms for detecting interword boundaries will allow users to speak without pauses between words

Speaker Verification

Although the V1000 has already demonstrated its ability to support speaker verification and identification applications, even in the word recognition mode, additional performance can be achieved with special purpose algorithms

All of the above features will be made available as upgrades to existing V1000 systems

SPECIFICATIONS

Audio Input

Low level: Dynamic microphone 20 mv rms
High level: 1 volt rms
Connectors: Both are ¼ inch phone jack

Digital Input/Output

RS-232C port, ASCII character code, up to 19.2 kbaud

Accuracy

99+% (measured by factory test tapes)

Noise Immunity

Maintains performance with background noise from conversations, machines and music up to 85 dB noise level

Vocabulary

2 second maximum duration of each word or utterance. 100 seconds word storage (approximately 160 typical words single-trained, 80 words double-trained)

Physical Dimensions

Width 11 inches
Depth 18 inches
Height 3½ inches

Electrical

115 volts ±10%, 50-60 Hz, 30 watts

Warranty

90 days for parts and labor

To arrange a convenient demonstration of the V1000 or to receive additional information on Votan's technology please call or write

VOTAN

26046 Eden Landing Road, Unit 7
Hayward, California 94545
(415) 785-8060

CHAPTER 5

OVERVIEW OF SPEECH SYNTHESIS PRODUCTS AND TECHNOLOGY

This chapter contains a brief review of currently available speech synthesis technology, plus a statement of Coast Guard operational requirements in this area. Three categories of subject matter discussed below are: 1) price ranges of speech synthesizers, 2) different product levels, and 3) Coast Guard operational requirements related to speech synthesis technology.

1) Price ranges of speech synthesizers. Speech synthesis products vary widely in price. For example, we note that Centigram's complete voice development system (model 6700) costs approximately $29,500. On the other end of the scale, we note various board and chip level products costing several hundred dollars, or less. Obviously, speech synthesizers vary widely in performance as a function of price. This report details the various advantages and disadvantages of each of the speech synthesis products reviewed, so that price value can be determined for any systems of potential interest to the Coast Guard.

2) Different product levels. This report notes that speech synthesis products are in three basic categories: LSI chip-level products, printed circuit board products, and complete synthesis systems.

LSI chip level products generally come with no control software and must be integrated into circuit boards before they can be used. This report cautions against using such products without being fully aware of the engineering and developmental costs which accompany such speech synthesizers.

Printed circuit board products are designed to plug into the interface units of existing host computers (RS232-C or parallel interfaces). Board-level synthesizers are generally easy to integrate into existing host computers. Elaborate software is generally not required, as synthesizers of this type generally operate under ASCII input, from a host terminal.

Complete speech synthesis systems are the most functional. For example, the Centigram 6700 Voiceware Development System is in this category. It comes complete with a host computer, CRT terminal, digitizer, and Centigram's Lisa synthesizer. This type of system requires nothing more than being plugged into a wall socket for operation. Such a system is very easy to operate, though it tends to be relatively expensive.

3) Coast Guard operational requirements related to speech synthesis technology. Overall the operational requirements for Coast Guard broadcasts are for a synthesizer with an essentially unlimited vocabulary. This would allow for broadcasting vessel names, geographical

locations, and detailed information. One application of speech synthesis technology that could be done with a relatively small vocabulary would be of synthesizing weather broadcasts which can be generated instantaneously from the teletype messages. Speech synthesis technology would also offer the option of being able to synthesize either female or male voices. For ease of operation by Coast Guard personnel, a text-to-speech voice synthesizer would be a very logical method for meeting Coast Guard requirements for speech synthesis. Such an approach merely requires that the user type in the desired text to a CRT terminal, with desired voice output being automatically derived from the text-to-speech unit. Several such text-to-speech units were reviewed. For example, we suggest the two Votrax text-to-speech units, plus Telesensory's latest text-to-speech prototype unit. We had the opportunity to evaluate Telesensory's text-to-speech unit over the telephone. Anyone wishing to hear a demonstration of this unit should call (415) 569-6257.

5.1 AVAILABLE SPEECH SYNTHESIZERS

There were a number of companies which SCRI was able to locate that manufacture voice output devices (speech synthesizers) of various types and configurations. First, there are the analysis synthesis synthesizers, which have stored coefficient values obtained from real speech. Second, there are the rule synthesizers, which model speech upon various parameters, combinations of which are used for actual synthesis. Finally, there are the synthesizers which rely upon direct digitization and playback of real speech.

For readers unfamiliar with linear prediction coding or LPC analysis, reference is made to Markel, Gray, and Wakita (1973). In this SCRL Monograph, the authors detail LPC analysis which involves predicting data from past data samples. Such an approach is intimately related to multiple regression and to setting up digital filter coefficients which allow for an economical (in terms of bit rate) representation of speech data. LPC analysis is commonly used in analysis synthesis.

As with speech recognition products, speech synthesis products may be broadly divided into three general product classifications: systems level products, board level products, and chip level products. The systems level prolucts are generally characterizable by including a host computer and consisting of stand alone terminals which synthesize speech.

SCRL received information from the following twelve manufacturers of speech synthesizers: 1)Centigram, 2)General Instruments, 3)Interstate Electronics, 4) Kurzweil Computer Prolucts, 5) Maryland Computer Services, 6) Mimic, 7) MSC, 8) National Semiconductor, 9) Percom Data Co., 10) Telesensory Speech Systems, 11) Texas Instruments, and 12) Votrax.

Centigram

Centigram's Lisa synthesizer is in the category of board-level products, as it is designed to connect to an existing computer interface.

Centigram has recently introduced their Lisa speech synthesizer which uses parametric waveform coding intervals of 50 msec. This unit features a low bit rate and connects to an RS232-C interface. The Lisa has memory storage for 30-120 seconds of stored speech data.

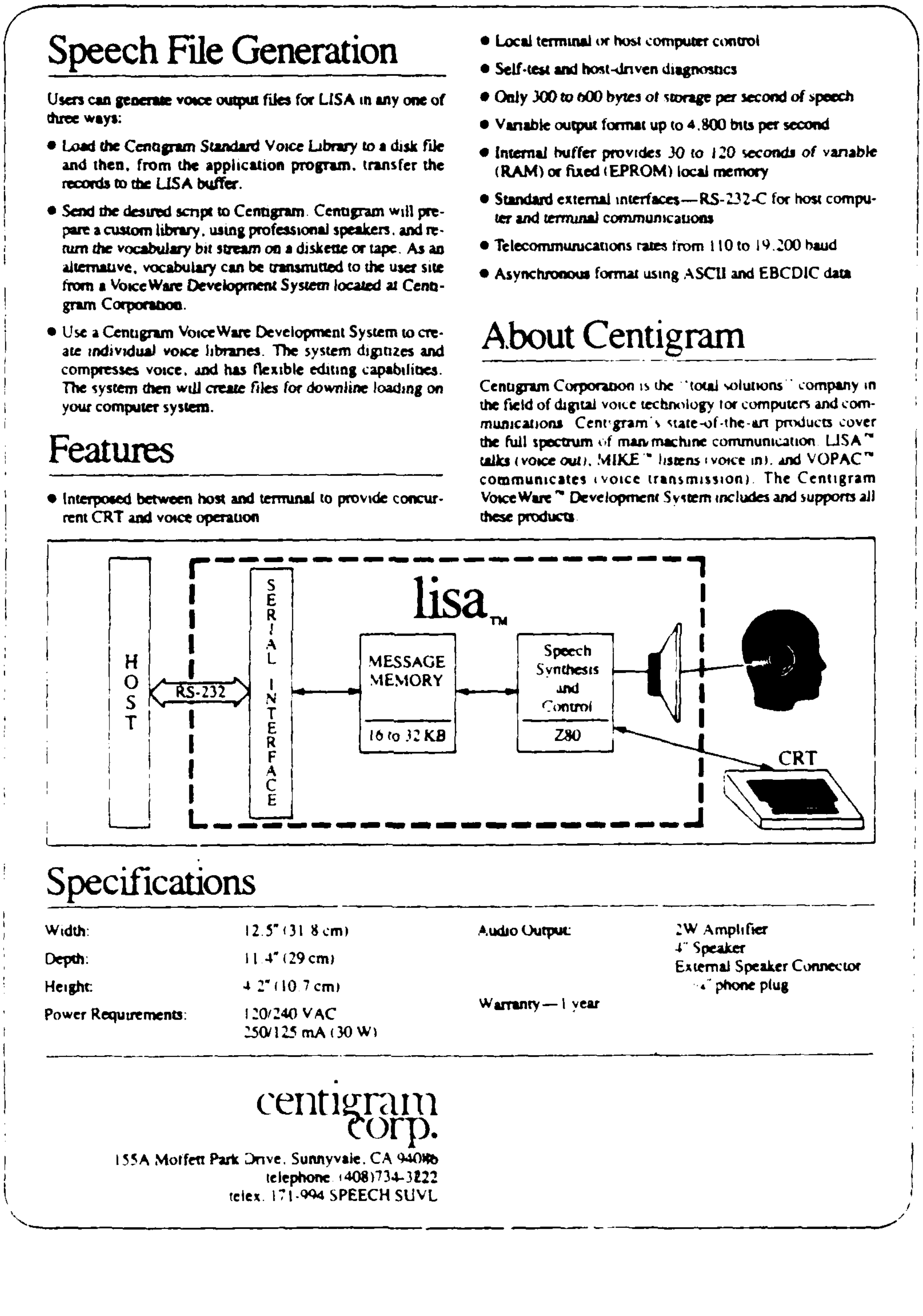
Centigram notes that their parametric waveform coding allows the user to reprogram the unit in the field, so that new utterances may be immediately stored for playback. This circumvents the problem encountered by most available synthesizers, which require reprogramming for synthesis of additional speech data at the manufacturer's base of operation. Table 21 indicates the specifications for the Lisa synthesizer which sells for $3,450.

The price is very reasonable considering the fact that the unit may be reprogrammed in the field to synthesize any desired utterance. Centigram emphasizes the high-quality voice output of the synthesizer.

Centigram also markets a large voiceware development system for $29,500 (model 6700). This system includes a digitizer, Lisa synthesizer, microcomputer, disk, floppy disk, and required software.

TABLE 21

Centigram's Lisa Parametric Waveform Coding Synthesizer



Speech File Generation

Users can generate voice output files for LISA in any one of three ways:

- Load the Centigram Standard Voice Library to a disk file and then, from the application program, transfer the records to the LISA buffer.
- Send the desired script to Centigram. Centigram will prepare a custom library, using professional speakers, and return the vocabulary bit stream on a diskette or tape. As an alternative, vocabulary can be transmitted to the user site from a VoiceWare Development System located at Centigram Corporation.
- Use a Centigram VoiceWare Development System to create individual voice libraries. The system digitizes and compresses voice, and has flexible editing capabilities. The system then will create files for downline loading on your computer system.

Features

- Interposed between host and terminal to provide concurrent CRT and voice operation
- Local terminal or host computer control
- Self-test and host-driven diagnostics
- Only 300 to 600 bytes of storage per second of speech
- Variable output format up to 4,800 bits per second
- Internal buffer provides 30 to 120 seconds of variable (RAM) or fixed (EPROM) local memory
- Standard external interfaces—RS-232-C for host computer and terminal communications
- Telecommunications rates from 110 to 19,200 baud
- Asynchronous format using ASCII and EBCDIC data

About Centigram

Centigram Corporation is the "total solutions" company in the field of digital voice technology for computers and communications. Centigram's state-of-the-art products cover the full spectrum of man/machine communication. LISA™ talks (voice out), MIKE™ listens (voice in), and VOPAC™ communicates (voice transmission). The Centigram VoiceWare™ Development System includes and supports all these products.

Specifications

Width:	12.5" (31.8 cm)	Audio Output:	2W Amplifier 4" Speaker External Speaker Connector ¼" phone plug
Depth:	11.4" (29 cm)		
Height:	4.2" (10.7 cm)		
Power Requirements:	120/240 VAC 250/125 mA (30 W)	Warranty — 1 year	

centigram corp.

155A Moffett Park Drive, Sunnyvale, CA 94086
telephone (408)734-3222
telex 171-994 SPEECH SUVL

General Instruments

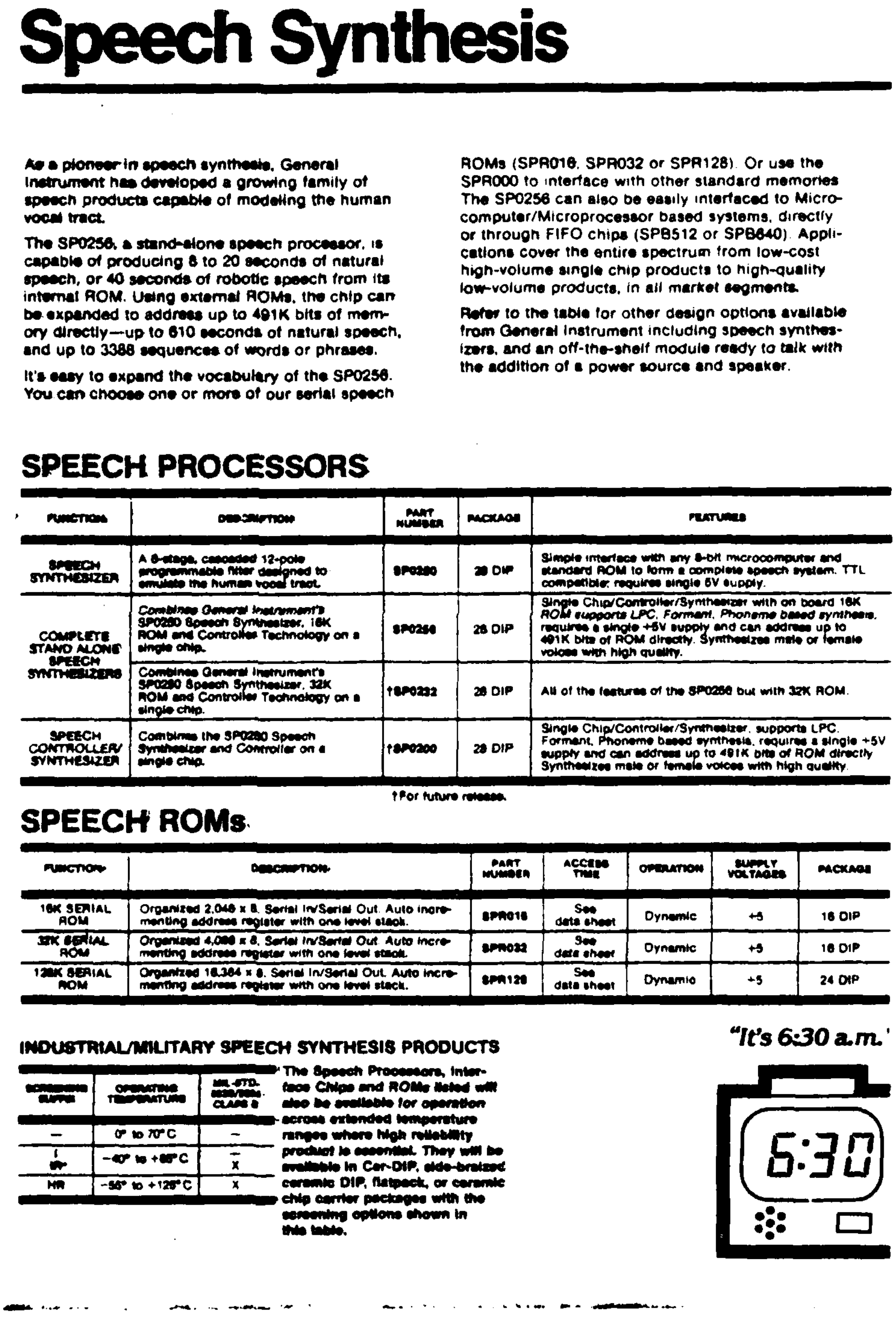
General Instruments' speech synthesis products are geared toward the chip level market. Again, this product classification involves sales to original equipment manufacturers who wish to use speech synthesis in existing products under development. General Instruments' basic approach to speech synthesis supports phoneme synthesis, even though it is an analysis synthesis approach.

General Instruments' basic LSI chip synthesizer is designated the SP0250. This chip contains circuitry for a 6-stage, cascaded 12-pole programmable filter designed to emulate the human vocal tract. The unit features simple interfacing with any 8-bit microcomputer and a standard ROM to form a complete speech system. The SP0250 chip is also used in General Instruments' stand alone speech synthesizers. First, the SP0250 is available with a 16K ROM and controller technology on a single chip, as the SP0256. Or, it is available with a 32K ROM and controller technology on a single chip as the SP0232 (for future release). Table 22 describes General Instruments' approach to speech synthesis, plus their speech processors and speech ROMs.

General Instruments markets several speech interface chips, plus a complete speech synthesis module. Their speech synthesis module VSM2032 combines the SP0250 synthesizer chip, a PIX1650A microcomputer (for formatting

TABLE 22

General Instrument Speech Processors and Speech ROMs



Speech Synthesis

As a pioneer in speech synthesis, General Instrument has developed a growing family of speech products capable of modeling the human vocal tract.

The SP0256, a stand-alone speech processor, is capable of producing 8 to 20 seconds of natural speech, or 40 seconds of robotic speech from its internal ROM. Using external ROMs, the chip can be expanded to address up to 491K bits of memory directly—up to 610 seconds of natural speech, and up to 3388 sequences of words or phrases.

It's easy to expand the vocabulary of the SP0256. You can choose one or more of our serial speech ROMs (SPR016, SPR032 or SPR128). Or use the SPR000 to interface with other standard memories. The SP0256 can also be easily interfaced to Microcomputer/Microprocessor based systems, directly or through FIFO chips (SPB512 or SPB640). Applications cover the entire spectrum from low-cost high-volume single chip products to high-quality low-volume products, in all market segments.

Refer to the table for other design options available from General Instrument including speech synthesizers, and an off-the-shelf module ready to talk with the addition of a power source and speaker.

SPEECH PROCESSORS

FUNCTION	DESCRIPTION	PART NUMBER	PACKAGE	FEATURES
SPEECH SYNTHESIZER	A 8-stage, cascaded 12-pole programmable filter designed to emulate the human vocal tract.	SP0250	28 DIP	Simple interface with any 8-bit microcomputer and standard ROM to form a complete speech system. TTL compatible; requires single 5V supply.
COMPLETE STAND ALONE SPEECH SYNTHESIZERS	*Combines General Instrument's* SP0250 Speech Synthesizer, 16K ROM and Controller Technology on a single chip.	SP0256	28 DIP	Single Chip/Controller/Synthesizer with on board 16K *ROM supports LPC, Formant, Phoneme based synthesis,* requires a single +5V supply and can address up to 491K bits of ROM directly. Synthesizes male or female voices with high quality.
	Combines General Instrument's SP0250 Speech Synthesizer, 32K ROM and Controller Technology on a single chip.	†SP0232	28 DIP	All of the features of the SP0256 but with 32K ROM.
SPEECH CONTROLLER/ SYNTHESIZER	Combines the SP0250 Speech Synthesizer and Controller on a single chip.	†SP0200	28 DIP	Single Chip/Controller/Synthesizer, supports LPC, Formant, Phoneme based synthesis, requires a single +5V supply and can address up to 491K bits of ROM directly Synthesizes male or female voices with high quality.

†For future release.

SPEECH ROMs

FUNCTION	DESCRIPTION	PART NUMBER	ACCESS TIME	OPERATION	SUPPLY VOLTAGES	PACKAGE
16K SERIAL ROM	Organized 2,048 x 8. Serial In/Serial Out. Auto incrementing address register with one level stack.	SPR016	See data sheet	Dynamic	+5	16 DIP
32K SERIAL ROM	*Organized 4,096 x 8. Serial In/Serial Out. Auto incrementing address register with one level stack.*	SPR032	See data sheet	Dynamic	+5	16 DIP
128K SERIAL ROM	Organized 16,384 x 8. Serial In/Serial Out. Auto incrementing address register with one level stack.	SPR128	See data sheet	Dynamic	+5	24 DIP

INDUSTRIAL/MILITARY SPEECH SYNTHESIS PRODUCTS

SCREENING SUFFIX	OPERATING TEMPERATURE	MIL-STD-883B/38510 CLASS B
—	0° to 70°C	—
I MY	−40° to +85°C	— X
HR	−55° to +125°C	X

The Speech Processors, Interface Chips and ROMs listed will also be available for operation across extended temperature ranges where high reliability product is essential. They will be available in Cer-DIP, side-brazed ceramic DIP, flatpack, or ceramic chip carrier packages with the screening options shown in this table.

"It's 6:30 a.m.'

speech data), and an RO-3-9333 ROM (for storing speech data). The unit is all contained on one printed circuit board and is designed to function as a speech synthesis evaluation module, with built-in filter, amplifier, on-board calculator, and clock vocabulary of 32 words and syllables which can be concatenated. The vocabulary can be modified by using custom ROMs or EPROMs. Table 23 lists the specificaticns for General Instruments' speech interface chips and speech synthesis module.

Interstate Electronics

Interstate's speech synthesis products are in the board level product classification.

In the preceding chapter of this report, it was noted that Interstate is a major manufacturer of speech recogniticn products. In this section of the report, we will only mention the characteristics of their VTM150 voice response module.

The VTM150 is a single printed-circuit board capable of phoneme synthesis (rule synthesis) of isolated or connected speech. The board provides a standard fixed vocabulary of approximately 500 words, and a user-programmable vocabulary of approximately 1000 words. The following two tables, 24 and 25, describe specifications of the Interstate VTM150 voice response module.

TABLE 23

General Instrument Speech Interface Chips and Voice

GENERAL
INSTRUMENT

SPEECH INTERFACE CHIPS

FUNCTION	DESCRIPTION	PART NUMBER	PACKAGE	FEATURES
SPEECH FIFO BUFFER MEMORY AND CONTROL LOGIC	8 bit x 64 words FIFO buffer memory to provide speech data to SP0256 from sources other than General Instrument speech ROMs.	SPB612	40 DIP	Provides interface to address/control SP0256 from microprocessor based systems. In addition stores speech data for SP0256. Can also be used in conjunction with speech ROMs.
	10 bit x 64 words FIFO buffer memory to provide speech data to SP0256 from sources other than General Instrument speech ROMs.	SPB640	40 DIP	
INTERFACE CONTROL LOGIC	Serial to parallel conversion of address, parallel to serial conversion of data and other control logic.	SPR000	40 DIP	Enables SP0256 to access speech data from industry standard parallel memories.

VOICE SYNTHESIS MODULE

FUNCTION	DESCRIPTION	PART NUMBER	PACKAGE	FEATURES
COMPLETE SPEECH SYNTHESIS	The VSM2032 utilizes General Instrument's state-of-the-art technology to synthesize speech. It combines the SP0280 Digital Speech Synthesizer, a PIC1650A microcomputer to format the Speech Data and a RO-3-9333 ROM to store the Speech Data.	VSM2032	3.25" x 5.0" P C. board with 15 pin edge connector	Self sufficient speech synthesis evaluation unit, with built in filter and amplifier, on board calculator and clock vocabulary of 32 words and syllables can be concatenated to say over one billion phrases. The vocabulary can be modified by using custom ROM or EPROM.

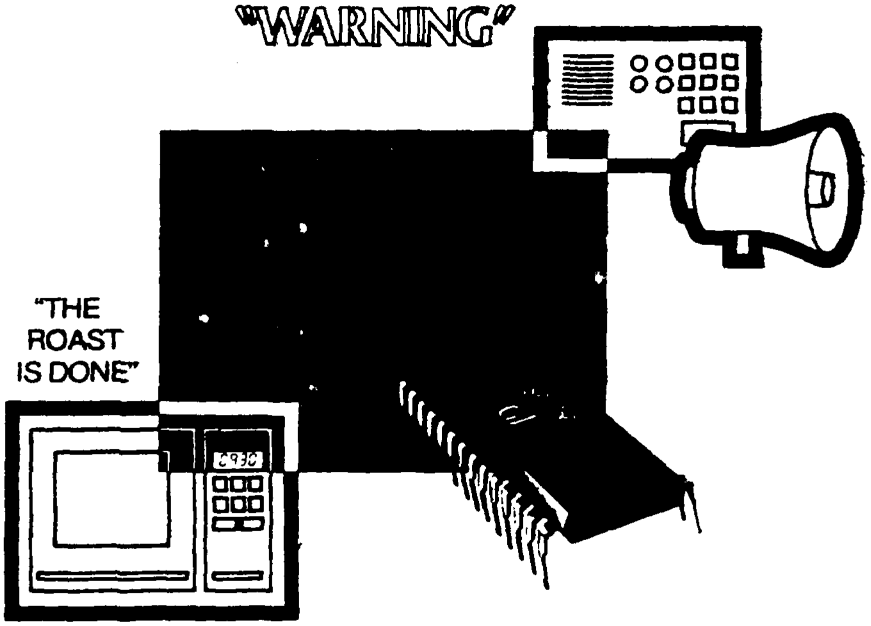
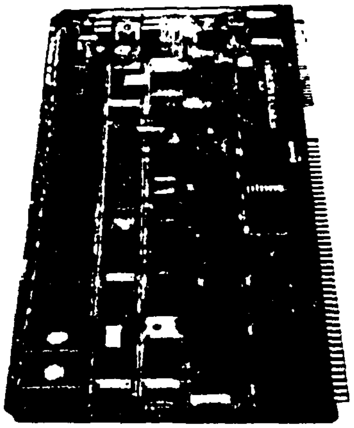


TABLE 24

Interstate VTM150 Voice Response Module

INTERSTATE ELECTRONICS CORPORATION

VOICE RESPONSE MODULE Model VTM150



Interstate's single-board Voice Response Module

- **Single-board voice response system**
- **500-word fixed vocabulary**
- **1000-word user programmable vocabulary**
- **High-quality synthetic speech**
- **Multibus™ parallel interface**
- **Serial and parallel ASCII communication ports**
- **2-watt output to external speaker**
- **Vocabulary generation, editing, and playback commands**

Single-Board Voice Response

The VTM150 is a single printed-circuit board capable of phonemic synthesis of isolated or connected speech with a low data rate from a controlling host processor. The board provides a standard fixed vocabulary of approximately 500 words and a user programmable vocabulary of approximately 1000 words.

The VTM150 is controlled via 12 commands: four commands for various playback functions with and without editing, four commands to control downloading with and without editing, one command to allow uploading all or specific vocabulary items, and three utility/system control commands.

Voice Response Module VTM150 contains a serial and a parallel port for host communication via ASCII characters and a Multibus™ parallel interface also controlled via ASCII characters. Each of the 64 phonemes and 4 inflection levels for each phoneme are sent from the host to the VTM150 as two ASCII characters to generate the user defined programmable vocabulary. Any vocabulary items in the fixed or programmable memory may be randomly selected and output to a listener via an ASCII word number.

The VTM150 delivers 2 watts of audio output into a 16-ohm speaker.

User Configuration Control

The VTM150 board contains a microprocessor, 4K-bytes of program EPROM, 4K-bytes of EPROM for fixed vocabulary, 10K-bytes of static RAM for programmable vocabulary and word number index file, a parallel output port with speech synthesizer integrated circuit and power amplifier, a host serial and parallel port, and a Multibus interface.

User configuration control is provided via eight control lines to the parallel I/O port shared by the speech synthesizer. Two of these lines select the user's mode of communication to the host; three select the serial word format; two select the parallel handshaking options; and one selects the termination character. Configuration control may be accomplished by either external TTL logic levels or directly by onboard switches.

VTM150 SPECIFICATIONS

Performance

Vocabulary Size: Approximately 500 words, fixed; 1000 words, user programmable.

Host Commands

Playback Commands:

PL – Playback word or words, including repeat and delay features

PA – Append and playback

PI – Insert and playback

PM – Modify and playback

TABLE 25

Interstate VTM150 Voice Response Module Flow Chart

Edit/Programmable Vocabulary Commands:
A – Append (insert phoneme string)
I – Insert (insert phoneme string at specified word)
D – Delete (by word or words)
M – Modify (delete and insert new string and re-sequence)

Save and Utility Commands:
S – Save/upload
↑B – Clear entire programmable vocabulary*
F – Free (displays available RAM and last word number)
B – Bit set for intercom control via parallel I/O

*↑ = *Control key*

Digital Input/Output

- Parallel TTL input/output – 8 data input, 8 data output, and 4 control.
- Asynchronous serial, RS232-C or current loop, 50 to 19,200 baud (switch selectable).
- Multibus parallel data transfer – 8 bidirectional data lines, 8 additional interface lines, and 4 Multibus communication lines.

Audio Output

2 watts into a 16-ohm speaker with onboard audio level adjustment.

Mechanical

Card Size: 6.75 x 12.0 x 0.062 inches (standard Intel Multibus card size)

Connector

Power: 86-pin, 0.156-inch spacing, Viking 2VH43/1AN or equivalent.

Signals: 60-pin, 0.100-inch spacing AMP PE5 14559 connector.

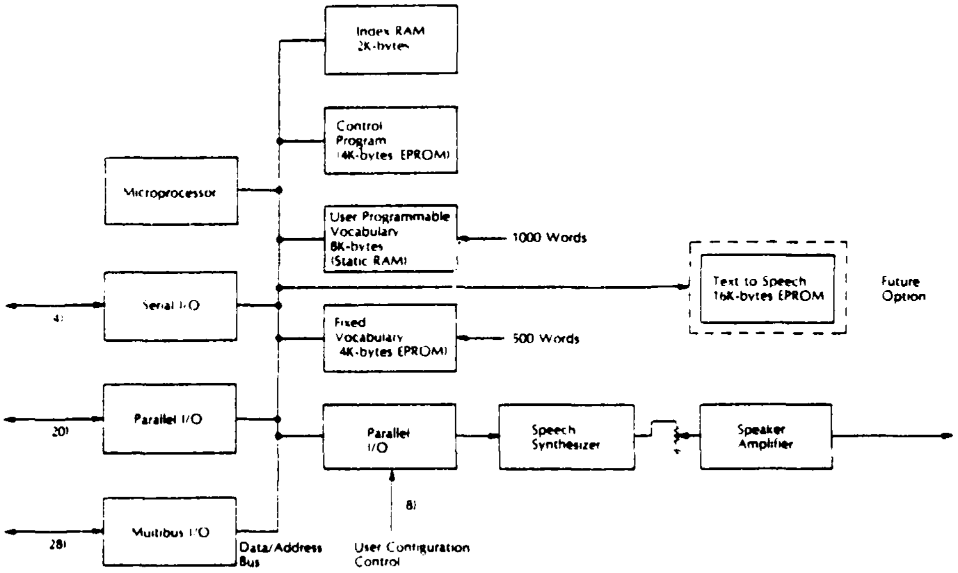
Electrical

Power Requirements: 590 mA at +5 Vdc; 100 mA at +12 Vdc; 110 mA at -12 Vdc.

Environmental

Temperature: 0 to 50°C.

Multibus™ is a trademark of the Intel Corporation.



Block Diagram of Voice Response Module VTM150

**INTERSTATE
ELECTRONICS
CORPORATION**

Voice Products Operations
1001 E. Ball Road, P.O. Box 3117, Anaheim, California 92803
Telephone 714/635-7210 TWX 910-591-1197 Telex 655443 & 655419
Call toll-free: in the continental U.S. 800/854-6979; in California 800/422-4580

Kurzweil Computer Products

Kurzweil Computer Products manufactures speech synthesis products which may be characterized as systems level products.

Kurzweil also manufactures products for the blind which include speech synthesis. In particular, note their Kurzweil Reading Machine Model III. Table 26 gives a short description of the unit.

Speech synthesis as an aid to the blind has been one of the earliest applications in mind for voice or speech generation devices. Kurzweil has been one of the leaders in considering the needs of the blind.

Maryland Computer Services

Maryland Computer Services produces speech synthesis terminals, designed to interface to an existing computer system. Thus, their products are between purely board level products and total systems level products, which generally include a host computer.

Maryland Computer Services does not actually manufacture speech synthesizers, but includes existing devices in their products which are basically oriented toward the blind. One of their products is the Total Talk computer terminal, which lists for $5,995. This unit uses a Votrax VSB synthesizer board which is a phoneme or rule synthesizer with 64 phonemes. It has a list of approximately 400 pronunciation rules.

TABLE 26

Kurzweil Reading Machine Model III

SPEECH OUTPUT TERMINAL CAPABILITY ADDED TO KURZWEIL READING MACHINE

The Kurzweil Reading Machine Model III now incorporates a Speech Output System that allows the Reading Machine to function as a full-word speech output device when connected to a suitable computer or computer terminal. This system permits the Reading Machine to be used as a receive-only terminal, analogous to a computer terminal. Although the Speech Output System does not entirely replace standard send-and-receive computer terminals, it can be used in conjunction with most ordinary terminals to produce speech output in place of print-outs.

The Speech Output System will accept ASCII Text presented through an RS-232 interface which is located at the rear of the Electronic Control Unit of the Reading Machine. The ASCII Text is stored in a 2000 character buffer, converted to phonemes and synthetically spoken. A complete set of keyboard instructions allows the user to back up in memory, repeat previously spoken lines or words, spell words and analyze punctuation.

The complete Speech Output System is contained in the digital cassette which also contains the standard Reading Machine System. The Reading Machine may be converted into a Speech Output System by means of a special command at the keyboard. While the RS-232 port is set to operate at 4800 Baud, the company will modify it on request to accept any Baud rate from 50 to 19200.

This computer voice output capability should open up new vocational possibilities for the visually handicapped in such places as data processing departments, financial institutions, reservations offices, and customer service departments, in which the ability to read such computer information is a must. It will also greatly facilitate research efforts of blind students, scientists, lawyers, and other professional who need access to computer information.

The terminal can switch from full words to spelled speech, and includes an adjustable speech rate, pitch, tone, and volume controls. The terminal can be set to automatically speak information going to or from the terminal. The unit also includes a speaking cursor key. Following is Table 27 which lists the Total Talk's specifications.

Maryland Computer Services also manufactures a number of other systems for use by the blind which incorporate speech synthesis. These include a talking telephone directory, a talking information management system, an automatic form writer, a talking word processing system, and a talking CRT terminal. These products illustrate the increasing use of speech synthesis in commercial products, in an application where it is of special benefit to blind users.

Mimic

Mi ic is a manufacturer of board level speech synthesis products.

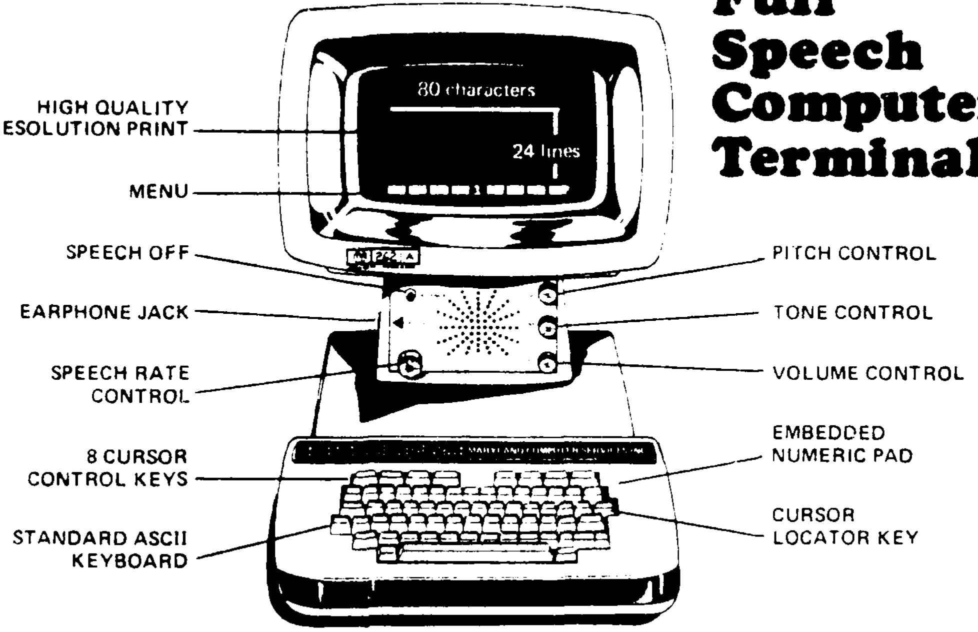
The Mimic speech processor is designed to synthesize speech on smaller computer systems, such as the TRS-80, Apple II, HP-85, etc. The unit consists of a board which digitizes speech into a bit stream which can be sampled by a computer, stored, and played back. The unit consists of an analog-to-digital converter and a digital-to-analog

TABLE [illegible]

Maryland Computer Services Total Talk Terminal

TOTAL TALK

Full Speech Computer Terminal



Specifications

Baud Rates – 110, 150, 200, 300, 600, 1200, 1800, 2400, 3600, 4800, 9600 and external

Asynchronous Interface – EIA Standard RS 232C (fully compatible with Bell 103A modems).

Transmission Modes – Full and Half duplex Asynchronous

Operating Modes – On Line, Off Line, Character Line.

Parity – Selectable Even, Odd, Zero, One.

Screen Capacity – 24 lines x 80 columns (1,920 characters).

Display Memory – 48 lines x 80 columns (3,840 characters)

8 Cursor Control Keys / Numeric Pad

Cursor Locator Key / ASCII Code Keyboard

Selectable Tabs and Margins

Full Editing Capabilities

Delivery – 90 Days

TOTAL TALK easily connects to most computer systems either directly or over a telephone line. The communication parameters are set from the keyboard and handle a wide range of protocols. All parameters can be vocalized, enabling the blind operator to change and verify them.

TOTAL TALK's many features make its use uncomplicated and straight forward. Tabs and margins are easily set. The cursor locator key vocally informs the operator of how many characters from the left margin and how many lines down from the top of the CRT screen that the cursor is positioned. Standard editing capabilities include inserting lines and characters, deleting lines and characters, clearing the entire display and underlining. A numeric data entry pad is embedded in the standard keyboard for easy entering of numbers.

For more information contact:

MARYLAND
COMPUTER
SERVICES INC

2010 Rock Spring Road
Forest Hill, Maryland 21050
(301) 838-8888 / 879-3366

converter (or a similarly-operating codec), with appropriate downsampling.

The Mimic speech processor is described as having a data rate of 9600 bits/second, which is a relatively high bit rate. Mimic notes that a 400-word vocabulary can be stored on one side of an 8-inch floppy disk (with an average word duration of .5 seconds).

The system comes whole, or in parts. A fully assembled and tested module costs $79 and a kit for the module costs only $19.95. Following is Table 28 which lists available Mimic units.

Mimic's speech processor unit is most interesting, and the $19.95 board kit has to be considered an unqualified bargain. The unit appears to have wide applications for experimenters interested in digital sampling and playback of voice. The unit could also apparently be used in conjunction with a host computer for speech recognition, given appropriate processing algorithms.

MSC

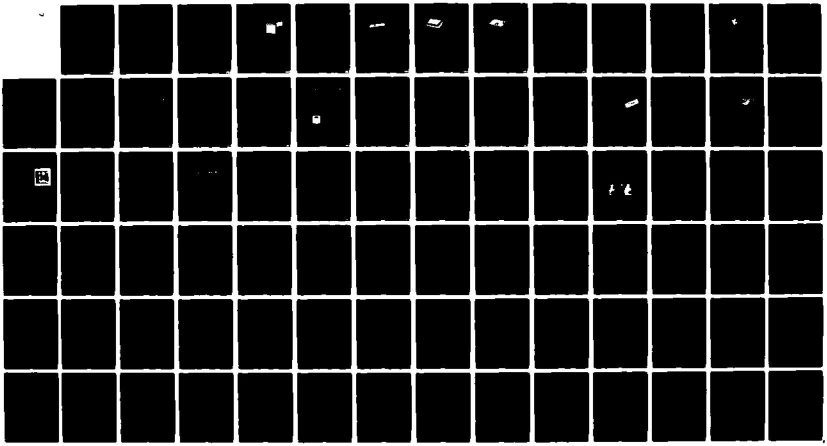
MSC's voice output products are board level products which are generally designed for purely commercial applications.

MSC manufactures voice output products which use LSI circuitry for actual recording of input speech data for subsequent playback. Thus, their devices are not truly

AD-A138 976 PRELIMINARY ANALYSIS OF AUTOMATIC SPEECH RECOGNITION AND SYNTHESIS TECHNOLOGY(U) SPEECH COMMUNICATIONS RESEARCH LAB LOS ANGELES CA J SHOUP ET AL. MAY 83

2/2

UNCLASSIFIED USCG-D-20-83 MDA904-82-C-0415 F/G 17/2 NL



END

DATE FILMED

[illegible]-84

DTIC

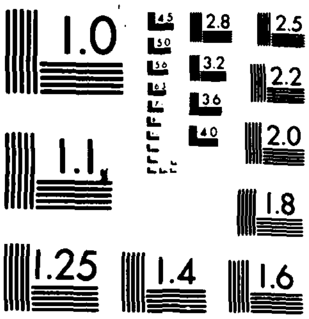

1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

TABLE 28

Mimic's List of Products

---

1. User's Manual for the MIMIC Speech Processor: $5. Contains complete theory of operation, schematics, assembly drawings, and S-100 Bus interface example with Z-80 (8080) driver program.
2. MIMIC audio demonstration cassette tape: $7.50. Compares MIMIC with other techniques in side-by-side listening tests.
3. Bare one-sided printed circuit board for the MIMIC Speech Processor: $19.95. Build it yourself. Manual not included.
4. MIMIC Speech Processor: $79 ($75 without manual). A fully assembled and tested module.
5. MIMIC System for Radio Shack's TRS-80: $169. Within minutes, you'll be demonstrating speech I/O on your computer. Table or wall mount. System includes manual, microphone, speaker with volume control, power supply, and a special cable assembly. Plugs into printer port on expansion interface, or use Radio Shack's Printer Interface Cable #26-1411 to connect to bus.
6. MIMIC System for Cromemco's TU-ART: $169. Similar to item #5 above, but with a different cable assembly.
7. MIMIC System for Parallel Port: $149. Can be wired directly to TTL port on most computers. Similar to item #5 above, but uses a standard DIP jumper instead of a special cable assembly.

** Available soon: MIMIC Systems for ZX-80, Apple, H-8, and HP-85. Let us know your interests, and we'll put you on our mail list.

** Note: For all MIMIC Systems, deduct $4 from list price if a manual is not required, and $10 if power supply not required.

8. S-100 Bus wire-wrap MIMIC interface card: $79. As described in manual, fully assembled. Large area for additional user logic. MIMIC System for Parallel Port, without power supply, plugs directly into this card (order items #7&8 for $218 total).
9. STD Bus wire-wrap MIMIC interface card: $79. Similar to #8.

speech synthesizers in the strict definition of the term. Yet, as their devices perform nearly identical functions as do comparable rule or analysis synthesizers, they have been included in this section dealing with voice output devices. MSC notes that there are a number of advantages to their approach as compared to other speech synthesis techniques. First, their devices claim excellent voice quality, which is, "indistinguishable from live voice". Certainly, this cannot be said of most commercial speech synthesizers. MSC also notes that their approach uses no moving parts, as do analog tape transports. Similarly, MSC's LSI circuitry avoids audio degradation associated with repeated playback of audio tapes.

MSC's top-of-the-line device is the 1650 Programmable Voice Readout System (VRS). The modular design of the 1650 VRS can accommodate 10 plug-in circuit boards, each with a capacity of 16 words stored in fragments of 406 milliseconds on individual ROMs and PROMs. Thus the vocabulary can be expanded to 160 words of the user's choice. MSC will custom build 1650 systems to include the words specified by the customer. Table 29 describes the 1650 VRS's specifications. Note that the 1650 comes complete with ROMs that are preprogrammed with standard MSC words, plus programmable ROMs ready to accept words of the user's choice. The 1650 has a list price of $650 and vocabulary is $50 per digit.

TABLE 29

MSC 1650 Programmable Voice Readout System

1650

Programmable Voice Readout System (VRS)

- Solid-state reliability
- Wide variety of applications
- High-fidelity voice duplication
- Expandable to 160 spoken words of your choice

The Model 1650 lets you add custom words to its standard vocabulary without paying custom charges.

The system comes complete with Read Only Memories (ROMs) that have been pre-programmed with standard MSC words, plus Programmable Read Only Memories (PROMs) ready to accept words of your choosing.

Its modular design accommodates 10 plug-in circuit boards. Each board has a capacity of 16 words which are stored in fragments of 406 milliseconds on individual ROMs and PROMs. A vocabulary can be expanded to 160 words within the standard ½ ATR rack. The desired message, which can be accessed instantly whenever needed, is "spoken" with such quality and clarity it's indistinguishable from a live announcement.

This proven, binary addressable system is currently being used throughout the world in critical applications such as aircraft warning systems, hospitals, refineries, chemical plants, and telecommunication and information systems of every kind.

The Model 1650, like all our solid-state readout systems, has no tapes to replace and no moving parts which could jam or wear; it operates virtually maintenance-free.

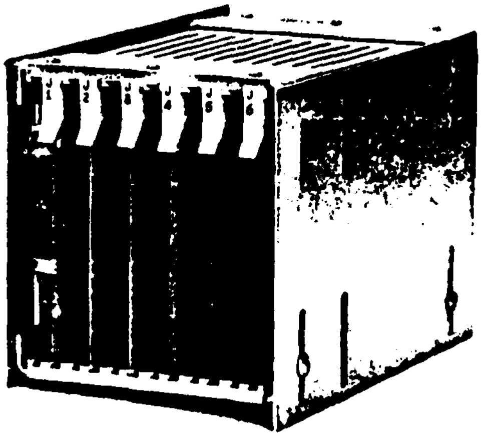
Specifications

Physical Size: ½ standard ATR rack (10.57" W × 5.2" H × 8.58" D)
Output, Audio: −6 dbm to 0 dbm balanced
Power Supply: +12 VDC
Input Power: 25 watts (max)
Operating Temperature: 0°C to 70°C
Input Format: Binary address
Output Transformer, isolated 600 ohms; dual audio circuit output provides 8 ohm @ 250 mw for monitoring

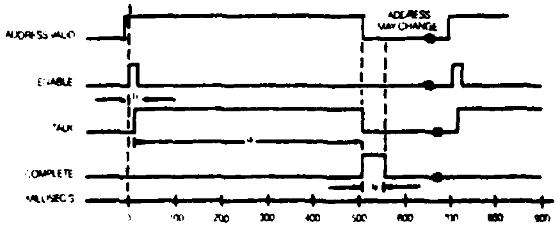
NOTE: Many standard interfaces are available. Please contact the factory for detailed information.

Ordering Information

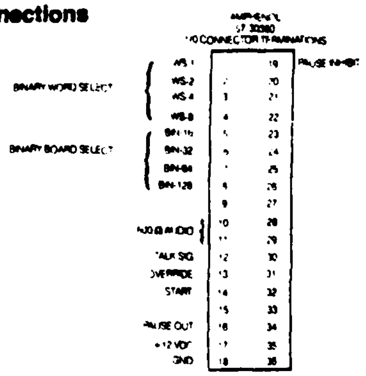
All orders for Model 1650 systems are custom built to include the words specified by the customer. MSC uses a Specification Sheet ordering system to control individual customer requirements and assigns a specific part number to each customer. Contact the factory for details on ordering information.



Timing Diagram for Binary Input



Output Connections



MSC offers the 1700 Voice Readout System (VRS). This unit features a similar approach to that of other MSC voice output products.

The 1700 VRS is designed for use where output of medium-length spoken messages are required. The one-board unit contains circuitry necessary to produce 16 words; a second circuit board may be added to expand the vocabulary to 32 words. Table 30 lists the 1700 VRS's specifications. The 1700 VRS has a list price of $650 (with 50 digits).

For situations requiring vocabulary changes, MSC recommends their 1750 VRS system, which stores individual words on programmable ROMs. Thus, any vocabulary can be specified without incurring setup or masking charges. Pause durations of the 1750 can be varied from 0-150 msec. The 1750 has a list price of $900 (with 10 words). Table 31 provides a description of the VRS 1750 specifications.

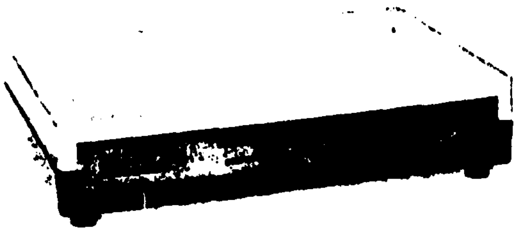
MSC also markets an automatic number announcer and an audio playback unit number announcer for use by telephone companies. For repeated broadcast of fixed messages, MSC markets the DCA-1 Dual Channel Annunciator described in Table 32. The unit has two channels for simultaneous output of a single message stored on programmable ROMs. The use of PROMs precludes charges for setup or masking. Standard message lengths are available up to five seconds, and the memory storage section can be expanded for longer durations.

TABLE 30

MSC 1700 Voice Readout System

1700

Voice Readout System (VRS)



- Solid-state reliability
- Wide variety of applications
- High-fidelity voice readout
- Up to 16 standard words on one circuit board
- Expandable to 32 words
- Packaged assembly available

The Model 1700 has become a standard in the telecommunications industry and it's finding new applications every day

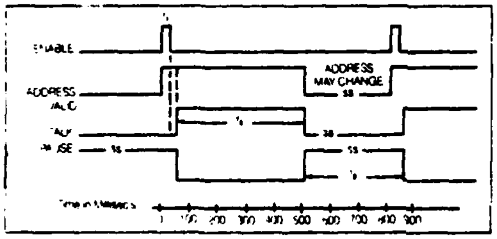
The unit is ideal for practically any situation where a medium-length spoken message is required, such as paging systems, computer and alarm systems, elevator floor announcements, credit card verifications, malfunction alerts and hotel/motel wake-up calls

The one-board unit contains all the circuitry necessary to produce 16 spoken words of extraordinary fidelity, and a second circuit board may be added to *expand its natural sounding vocabulary to 32 words*

Since all words are stored in separate Read Only Memories (ROMs), they can easily be added up in any sequence desired. The 16 standard spoken words are: "zero" through "nine," "plus," "minus," "times," "divide," "equal" and "point." Additional words of your choosing are subject to a one-time setup charge. The first 10 numeric words accept either binary address or 10 mutually exclusive switch closures. Additional words must utilize binary address

MSC's Model 1700 VRS is available as a circuit board only, or as an enclosed assembly.

Timing Diagram



Output Connections

OUTPUT PINS FOR LOW-COST VOICE SWITCH CLOSURE

CONNECTOR PINS	SIGNALS
1	AUD 1
2	VCC (+5V)
3	20KHZ or AD2048
4	Dig 1
5	8 Ω SPEAKER
6	Dig 0
7	Dig 2
8	Dig 3
9	Dig 7
10	Dig 6
11	Dig 5
12	Dig 4
13	ENABLE
14	OVRD*
15	TALK
16	Dig 9
17	Dig 8
18 19	+12V OR +6V
20	PAUSE
21	TONE
22	AUD 2
31	−12V or −6V
32	−6
33 34	GND

NOTE: * when override is High Logic, the system is disabled

OUTPUT PINS FOR LOW-COST VOICE BINARY INPUT

CONNECTOR PINS	SIGNALS
1	Bin 2
2	Bin 8
3	Bin 4
4	Bin 1
5	PAUSE
6	OVRD*
7	TONE
8	20KHZ
9 28	GND
10 27	VCC
11 14	+12V or +6V
13	8 Ω SPEAKER
[illegible]	BIN 16
15 31	−12V OR −6V
17	PS1
18	ENABLE
19	TALK
20	AUD 1
21	AUD 2
30	PS2

NOTE: * when override is Low Logic, the system is disabled

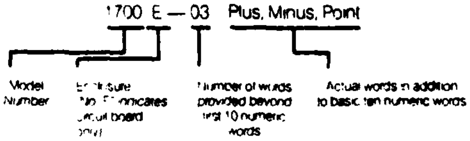
Model Numbers

1700E	+6 VDC 10 Switch Closures
1701E	$\substack{+12\\-12}$ and +5 VDC 10 Switch Closures
1702E	+6 VDC Binary input
1703E	$\substack{+12\\-12}$ and +5 VDC Binary input

Specifications

Physical Size: 8" W × 1¾" H × 5½" D
Output: Audio −6 dbm to 0 dbm
Power Supply: ±12 VDC and +5 VDC or #6 VDC
Input Power: 2.5 watts (max) for 10 words
Operating Temperature: 0°C to 70°C
Output Transformer: 600 balanced and 8 ohms 250 mw
Standard Interface: 34 pin 3m ribbon P/N 34 14-0000

Ordering Information



1700 E — 03 Plus, Minus, Point
Model Number
Enclosure (No E indicates circuit board only)
Number of words provided beyond first 10 numeric words
Actual words in addition to basic ten numeric words

The standard Model 1700's shown on the chart below include the 10 numeric words zero to nine. The number of additional words specified by the customer beyond the first 10 numeric words must be added to the model number as a dash number. Up to 22 additional words may be specified. If your system does not require the first 10 numeric words, consult factory for special model number. Any additional requirements not covered by these models may be ordered by consulting the factory for details.

TABLE 31

MSC 1750 Voice Response System

1750

Voice Response System

- Up to 32 custom words without custom charges

Like the Model 1700, this system features a 16-word spoken vocabulary expandable to 32 (using two circuit boards).

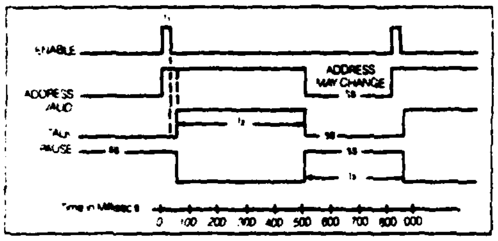
And it's similar to the Model 1700 in more ways than one. For example, it operates virtually maintenance-free because it has no tapes or moving parts of any kind. And its reliable, solid-state circuitry provides excellent, natural sounding voice reproduction that can scarcely be distinguished from the original.

The main difference is, the Model 1750 stores individual words on Programmable Read Only Memories (PROMs), instead of ROMs. That way, you can specify any vocabulary you like without incurring setup or masking charges.

So, for applications requiring vocabulary changes, the Model 1750 is a wise choice.

It accepts binary address only, and features a Pause Override Control that lets you adjust the duration of the pause from 0 to 150 microseconds.

Timing Diagram



Specifications

Physical Size: 8" W × 1¾" H × 5½" D
Output, Audio: −6 dbm to 0 dbm
Power Supply: ±12 VDC and +5 VDC or #6 VDC
Input Power: 2.5 watts (max) for 10 words
Operating Temperature: 0°C to 70°C
Output Transformer: 600 balanced and 8 ohms 250 mw
Standard Interface: 34 pin 3m ribbon P/N 34.14-0000

Output Connections

OUTPUT PINS FOR LOW-COST VOICE SWITCH CLOSURE

CONNECTOR PINS	SIGNALS
1	AUD 1
2	Vcc (+8V)
3	20KHZ or ADR048
4	Dig 1
5	8 Ω SPEAKER
6	Dig 0
7	Dig 2
8	Dig 3
9	Dig 7
10	Dig 6
11	Dig 5
12	Dig 4
13	ENABLE
14	OVRD*
15	TALK
16	Dig 9
17	Dig 8
18 19	+12V OR +8V
20	PAUSE
21	TONE
22	AUD 2
31	−12V or −8V
32	PS
33 34	GND

NOTE *When override is High Logic "1" the system is disabled

OUTPUT PINS FOR LOW-COST VOICE BINARY INPUT

CONNECTOR PINS	SIGNALS
1	Bin 2
2	Bin 8
3	Bin 4
4	Bin 1
5	PAUSE
6	OVRD*
7	TONE
8	20KHZ
9 29	GND
10 27	Vcc
11 14	+12V or +8V
13	8 Ω SPEAKER
15	Bin 16
16 33	−12V OR −8V
17	PS1
18	ENABLE
19	TALK
28	AUD1
30	AUD2
20	PS2

NOTE *When override is Low Logic "0" the system is disabled

Ordering Information

1700 E — 03 Plus, Minus, Point

Model Number

Enclosure (No "E" indicates circuit board only)

Number of words provided beyond first 10 numeric words

Actual words in addition to basic ten numeric words

Model Numbers

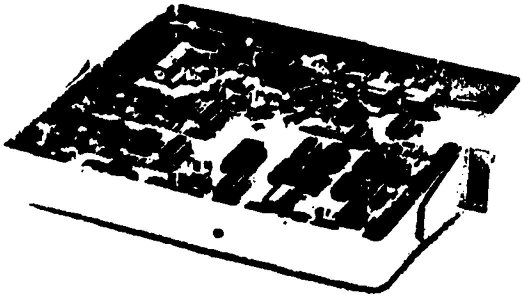
1700E	+6 VDC, 10 Switch Closures
1701E	$^{+12}_{-12}$ and +5 VDC, 10 Switch Closures
1702E	+6 VDC, Binary Input
1703E	$^{+12}_{-12}$ and +5 VDC, Binary Input

TABLE 3?

MSC DCA-1 Dual Channel Annunciator

DCA-1

Dual Channel Annunciator



- Solid-state reliability
- Completely automatic
- Simultaneous output
- Easy installation
- Low maintenance

For repeated broadcasting of fixed messages, MSC's reliable DCA-1 is hard to beat

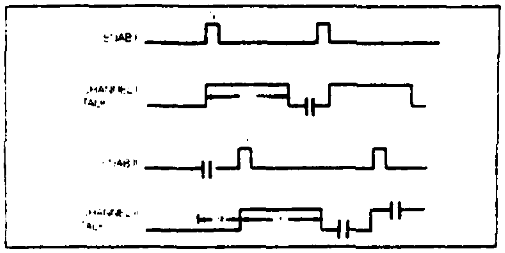
It features life-like voice reproduction with natural attenuation and spacing Messages flow smoothly and are easily understood

The DCA-1 provides two independent channels for simultaneous output of a single message stored on-board in Programmable Read Only Memories (PROMs) The use of PROMs precludes any charge for setup or masking

Standard message lengths are available up to 5 seconds, although the memory storage section may be expanded to accommodate longer durations

There are no recording tapes to stretch or break, and no moving parts to wear. Maintenance of this all-solid-state system is almost zero

Timing Diagram for DCA-1

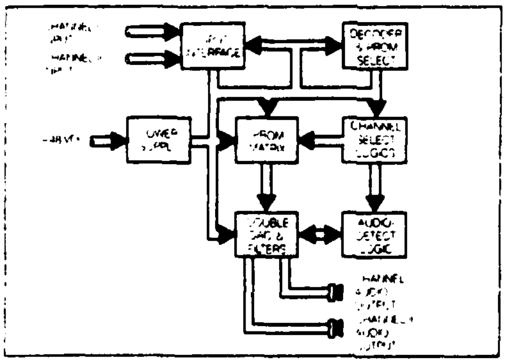


PRIMARY RECORDING LAST DURATION OF MESSAGE

VARIABLE DELAY

[illegible] OTHERWISE SPECIFIED

Block Diagram for DCA-1



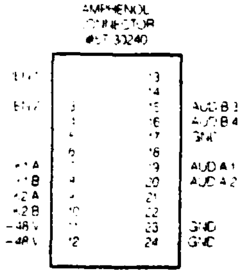
Specifications

Physical Size 10.5" W × 9.4" D × 1.5" H
Output, Audio −6 dBm + 0 dBm
Power Supply −48VDC (±5VDC regulated lamp. *±12VDC regulated 100 ma)*
Input Power 30 Watts
Input Format Switch closures
Operating Temperature 0°C to 70°C

Ordering Information

Consult factory

Output Connections



National Semiconductor

National Semiconductor markets speech synthesis chips which are for O.E.M. use. This involves incorporating speech synthesis chips into existing circuit boards. Rather than describing National's extensive line of LSI chip products related to voice output, this report focuses upon their basic synthesizer chips only. National's LSI chips are sold without any control software.

National Semiconductor notes that their digitalker chips will synthesize voices for males, females, or children. They market two basic synthesis chips.

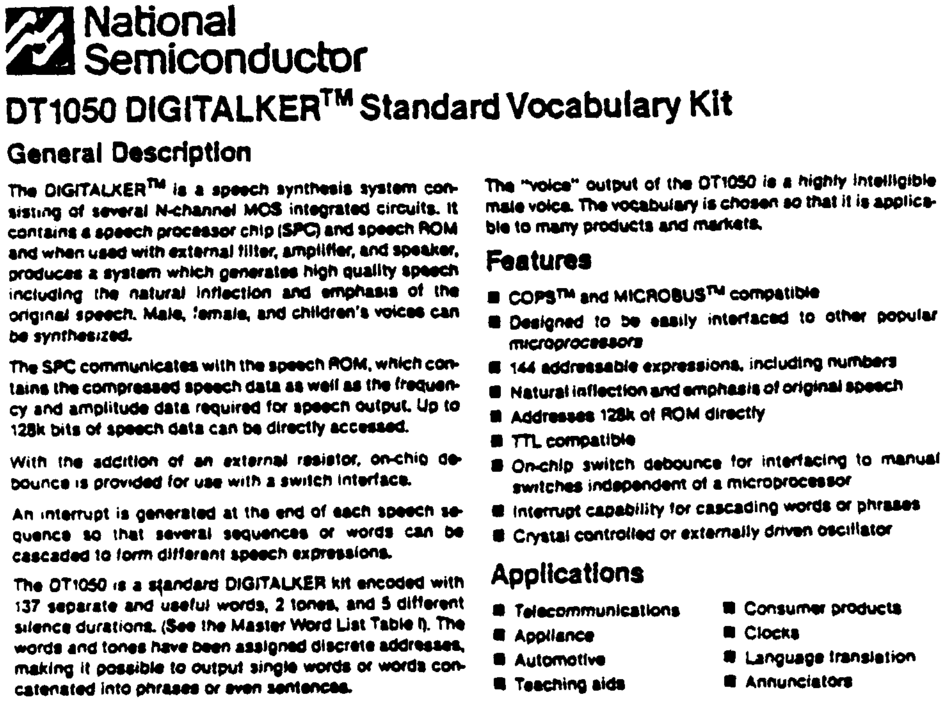
First, there is National's DT 1050 Digitalker kit. This chip is intended, generally, for O.E.M. applications (calculators, etc.). This kit features a chip with 137 words, two tones, and five pause durations. This kit sells for approximately $90. It can be used in conjunction with various computers, where the user can supply appropriate control software. Table 13 gives a block diagram of the DT1050 kit.

National also markets the DT 1000 Digitalker. This unit features a 138-word vocabulary, five silence durations, and a 1/2-watt on-board amplifier.

Note that SCRL did not receive any reply to letters sent to National Semiconductor inquiring about their speech products. Descriptions of National's speech synthesizers came from computer magazines, where their chips are commonly

TABLE 33

National Semiconductor DT 1050 Kit



National Semiconductor

DT1050 DIGITALKER^{TM} Standard Vocabulary Kit

General Description

The DIGITALKER^{TM} is a speech synthesis system consisting of several N-channel MOS integrated circuits. It contains a speech processor chip (SPC) and speech ROM and when used with external filter, amplifier, and speaker, produces a system which generates high quality speech including the natural inflection and emphasis of the original speech. Male, female, and children's voices can be synthesized.

The SPC communicates with the speech ROM, which contains the compressed speech data as well as the frequency and amplitude data required for speech output. Up to 128k bits of speech data can be directly accessed.

With the addition of an external resistor, on-chip debounce is provided for use with a switch interface.

An interrupt is generated at the end of each speech sequence so that several sequences of words can be cascaded to form different speech expressions.

The DT1050 is a standard DIGITALKER kit encoded with 137 separate and useful words, 2 tones, and 5 different silence durations. (See the Master Word List Table I). The words and tones have been assigned discrete addresses, making it possible to output single words or words concatenated into phrases or even sentences.

The "voice" output of the DT1050 is a highly intelligible male voice. The vocabulary is chosen so that it is applicable to many products and markets.

Features

- COPS^{TM} and MICROBUS^{TM} compatible
- Designed to be easily interfaced to other popular microprocessors
- 144 addressable expressions, including numbers
- Natural inflection and emphasis of original speech
- Addresses 128k of ROM directly
- TTL compatible
- On-chip switch debounce for interfacing to manual switches independent of a microprocessor
- Interrupt capability for cascading words or phrases
- Crystal controlled or externally driven oscillator

Applications

- Telecommunications
- Appliance
- Automotive
- Teaching aids
- Consumer products
- Clocks
- Language translation
- Annunciators

marketed by second parties. Their chips allow the user to concatenate phonemes. Although National's synthesizer chips output speech which is immediately distinguishable from live speech, they do so at a relatively low price.

Percom Data Co.

Percom's speech synthesis products are in the category of what might roughly be characterized as board level products. Their peripheral devices are designed to plug into smaller computer systems.

Percom markets a variety of peripheral devices for smaller computer systems, such as the TRS-80. One of these devices is a module designed to let users control LPC synthesized output from the Texas Instruments' (TI's) Speak & Spell unit. The unit uses a 9-volt battery or a standard calculator power pack. The unit requires Level II BASIC, a 4K memory, and an expansion interface or printer cable adaptor. Following is Table 34 on the Speak-2-Me-2 interface module which retails for $69.95.

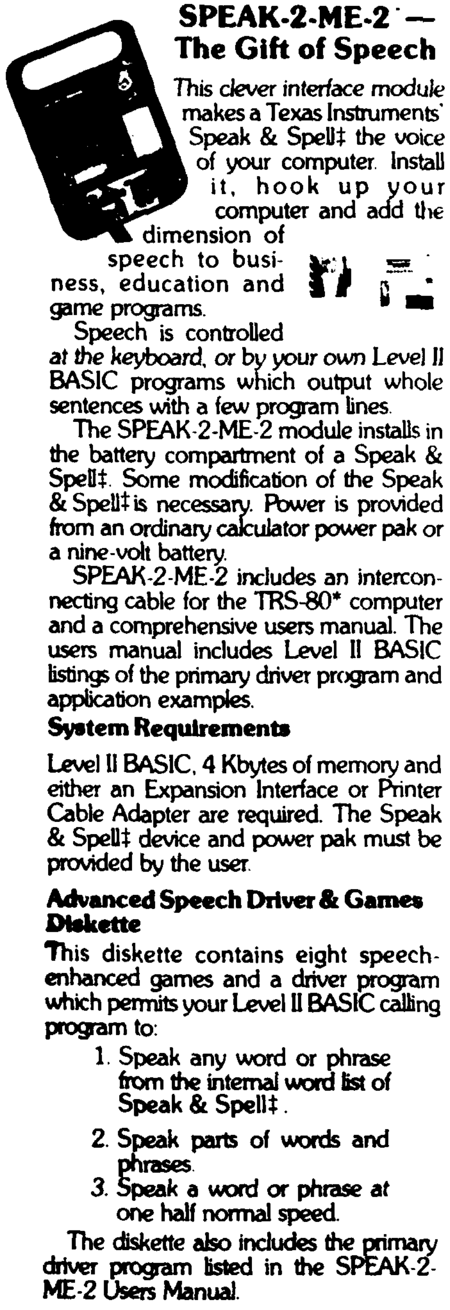
Telesensory Speech Systems

Telesensory's speech synthesis products are in the category of board level products.

Telesensory markets a number of synthesizers which use stored LPC coefficients for actual synthesis. The

TABLE 34

Percom Speak-2-Me-2 Interface Module



SPEAK-2-ME-2 — The Gift of Speech

This clever interface module makes a Texas Instruments' Speak & Spell‡ the voice of your computer. Install it, hook up your computer and add the dimension of speech to business, education and game programs.

Speech is controlled *at the keyboard, or by your own* Level II BASIC programs which output whole sentences with a few program lines.

The SPEAK-2-ME-2 module installs in the battery compartment of a Speak & Spell‡. Some modification of the Speak & Spell‡ is necessary. Power is provided from an ordinary calculator power pak or a nine-volt battery.

SPEAK-2-ME-2 includes an interconnecting cable for the TRS-80* computer and a comprehensive users manual. The users manual includes Level II BASIC listings of the primary driver program and application examples.

System Requirements

Level II BASIC, 4 Kbytes of memory and either an Expansion Interface or Printer Cable Adapter are required. The Speak & Spell‡ device and power pak must be provided by the user.

Advanced Speech Driver & Games Diskette

This diskette contains eight speech-enhanced games and a driver program which permits your Level II BASIC calling program to:

1. Speak any word or phrase from the internal word list of Speak & Spell‡.
2. Speak parts of words and phrases.
3. Speak a word or phrase at one half normal speed.

The diskette also includes the primary driver program listed in the SPEAK-2-ME-2 Users Manual.

synthesizers which will be reviewed in this report are the Speech 1000 LPC board, the Series III Speech Synthesizer Module, a prototype text-to-speech system, and the S2B and S2C synthetic speech boards.

The Telesensory 1000 LPC board is noted to have superior voice quality and the capacity for large vocabulary storage (up to 458K, typically 200-300 seconds). The unit also features a variety of common interface options, including the popular RS-232C. The unit features a number of variable parameters for synthesizing speech, such as a variable and programmable audio gain and output, speech speed control, and interword pause control.

Telesensory Speech Systems notes that the Speech 1000 board is applicable for all l nguages. For natural intonation, Telesensory suggests building sentences around phrases or other sentences. Following is Table 35 with the Telesensory's Speech 1000 board, including a block diagram of the system's configuration. The Speech 1000 board has a retail price of $1,200 for single units. It is Telesensory's top-of-the-line synthesizer.

Telesensory Speech Systems produces the Series III Speech Synthesizer Module. This unit is lower-priced than the Speech 1000 board, costing $295 to $395, depending upon options selected. The Series III Synthesizer can accommodate both custom and standard vocabularies in standard ROMs or EPROMs up to a capacity of 256 utterances,

TABLE 35

Telesensory Speech Systems 1000 Board

System Specifications

Synthesizer
Telesensory's PDSP (Programmable Digital Signal Processor) chip set implementing a 12 Pole Lattice Filter Structure

Speech Encoding
Linear Predictive Coding: 2200 bits per second of speech is standard, other encoding rates available

Vocabulary Capacity
Approximately 200 seconds of speech at 2200 bps encoding rate, up to 300 seconds at lower rates

The available time may be used to store any number of words, phrases or sentences

Vocabulary Memory
Total of 7 28-pin sockets
Available for ROM, EPROM or RAM
Total capacity of 458k bits of standard semiconductor memory

Interfaces
Multibus: I/O Slave
Serial Port: (RS232C) 300-9600 Baud (Jumper Select)
Parallel Port: (TTL) 8 bits (Data), 3 bits (Control)

Audio
2 Watts into 8 ohms
Low Pass Filter: f_c=4.8kHz @-6dB
Rolloff: 42dB/octave
Programmable Amplitude Level: 8 levels, 3dB/level
Programmable Speech Speed: 2X normal to ½X normal

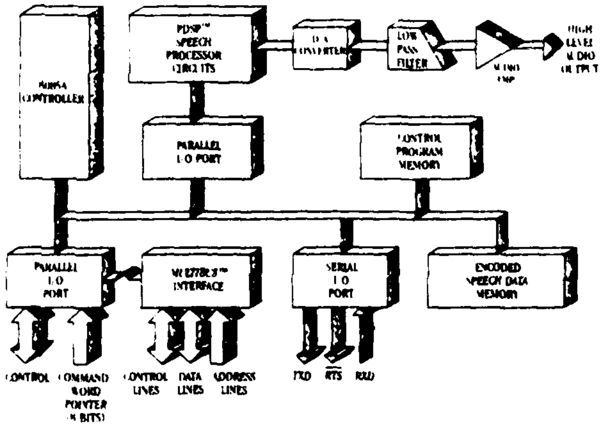
Power
+5V at 2 amps (max.)
+12V at 1 amps (max.)
−12V at 0.1 amps (max.)

Size
Intel's Multibus Board Form Factor
6.75" x 12.00" x 0.50" (17.15cm x 30.48cm x 1.27cm)

Weight
16 oz (454 gm)

Operating Temperature
0°C to 55°C

SPEECH 1000™ SYSTEM BLOCK DIAGRAM



TELESENSORY
Speech Systems
3408 Hillview Avenue • P.O. Box 10099
Palo Alto California 94304
(415) 856-8255 • Telex 348352

SPEECH 1000 and PDSP are trademarks of Telesensory Speech Systems
Multibus™ is a trademark of Intel Corporation

a division of
TELESENSORY
SYSTEMS,
INC.

Printed in U.S.A. 81-204 PART #547001 B

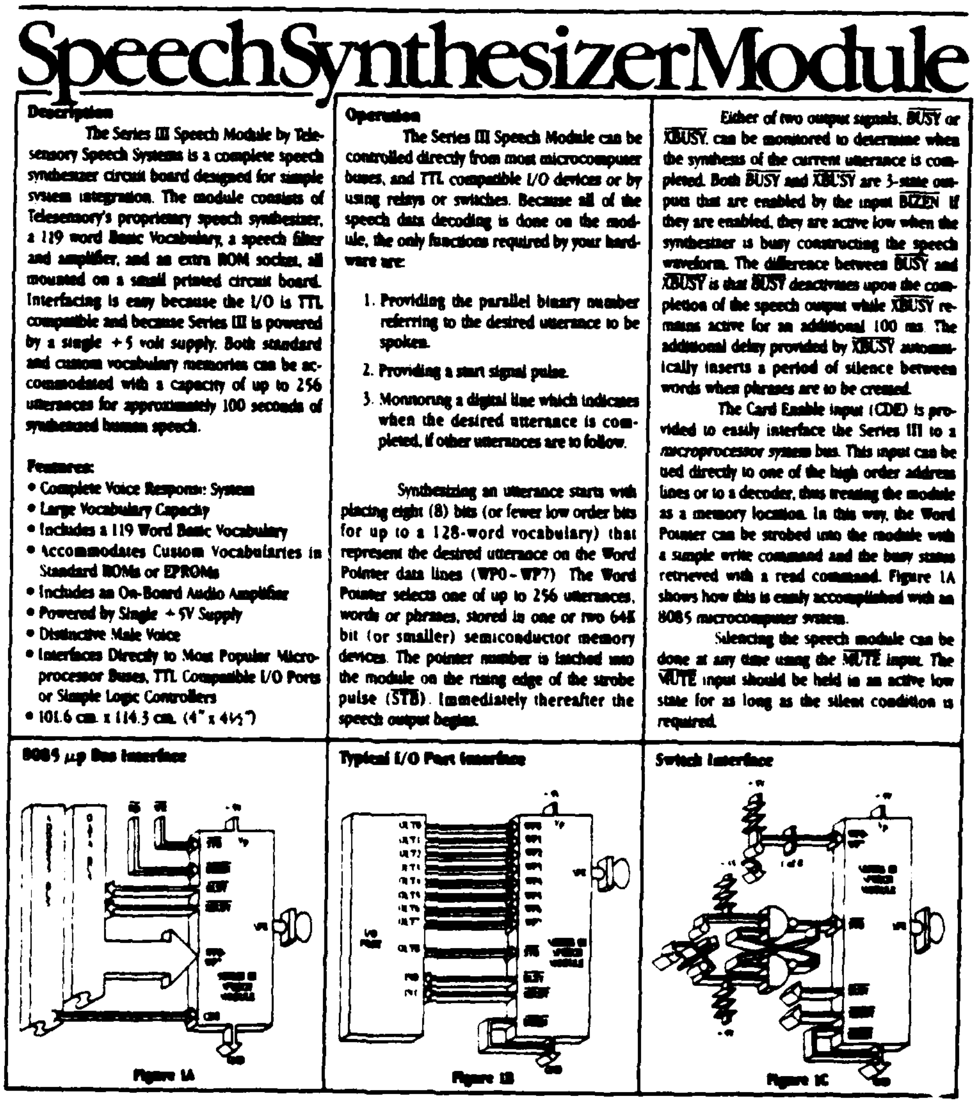
for a total time of approximately 100 seconds of synthesized speech. The unit is a complete voice response system, including an on-board audio amplifier. The unit interfaces directly with most popular buses, including TTL compatible I/O port, or simple logic controllers. The unit features a distinctive male voice, and has a relatively large vocabulary capacity. The unit is powered by a single +5V power supply. Following are Tables 36 and 37 covering the Series III Speech Synthesizer Module.

Telesensory Speech Systems notes that they are developing a prototype text-to-speech system, which will be a stand-alone unlimited speech peripheral device. The unit will feature an RS-232C interface. The text-to-speech system will include some prosodic features for sentences. Basically, the unit is described as having two modes: 1) lexical - for normal stress patterns, and 2) prosodic - where whole phrases are analyzed, and words are stressed in relation to surrounding words.

Telesensory markets two mini circuit boards for speech synthesis. The S2B and S2C boards feature the minimum components necessary for speech. The units include one or two 16K ROMs, depending upon vocabulary selection, plus clock frequency circuitry. Available vocabularies include a 24-word calculator vocabulary, and two 64-word general purpose vocabularies. The units are based upon the CRC synthesizer chip, which costs $65 with vocabularies running

TABLE 36

Telesensory Speech Systems Series III
Speech Synthesizer Module



SpeechSynthesizerModule

Description

The Series III Speech Module by Telesensory Speech Systems is a complete speech synthesizer circuit board designed for simple system integration. The module consists of Telesensory's proprietary speech synthesizer, a 119 word Basic Vocabulary, a speech filter and amplifier, and an extra ROM socket, all mounted on a small printed circuit board. Interfacing is easy because the I/O is TTL compatible and because Series III is powered by a single +5 volt supply. Both standard and custom vocabulary memories can be accommodated with a capacity of up to 256 utterances for approximately 100 seconds of synthesized human speech.

Features:

- Complete Voice Response System
- Large Vocabulary Capacity
- Includes a 119 Word Basic Vocabulary
- Accommodates Custom Vocabularies in Standard ROMs or EPROMs
- Includes an On-Board Audio Amplifier
- Powered by Single +5V Supply
- Distinctive Male Voice
- Interfaces Directly to Most Popular Microprocessor Buses, TTL Compatible I/O Ports or Simple Logic Controllers
- 101.6 cm. x 114.3 cm. (4" x 4½")

Operation

The Series III Speech Module can be controlled directly from most microcomputer buses, and TTL compatible I/O devices or by using relays or switches. Because all of the speech data decoding is done on the module, the only functions required by your hardware are:

1. Providing the parallel binary number referring to the desired utterance to be spoken.
2. Providing a start signal pulse.
3. Monitoring a digital line which indicates when the desired utterance is completed, if other utterances are to follow.

Synthesizing an utterance starts with placing eight (8) bits (or fewer low order bits for up to a 128-word vocabulary) that represent the desired utterance on the Word Pointer data lines (WP0 - WP7) The Word Pointer selects one of up to 256 utterances, words or phrases, stored in one or two 64K bit (or smaller) semiconductor memory devices. The pointer number is latched into the module on the rising edge of the strobe pulse ($\overline{STB}$). Immediately thereafter the speech output begins.

Either of two output signals, $\overline{BUSY}$ or $\overline{XBUSY}$, can be monitored to determine when the synthesis of the current utterance is completed. Both $\overline{BUSY}$ and $\overline{XBUSY}$ are 3-state outputs that are enabled by the input $\overline{BIZEN}$. If they are enabled, they are active low when the synthesizer is busy constructing the speech waveform. The difference between $\overline{BUSY}$ and $\overline{XBUSY}$ is that $\overline{BUSY}$ deactivates upon the completion of the speech output while $\overline{XBUSY}$ remains active for an additional 100 ms. The additional delay provided by $\overline{XBUSY}$ automatically inserts a period of silence between words when phrases are to be created.

The Card Enable input (CDE) is provided to easily interface the Series III to a microprocessor system bus. This input can be tied directly to one of the high order address lines or to a decoder, thus treating the module as a memory location. In this way, the Word Pointer can be strobed into the module with a simple write command and the busy status retrieved with a read command. Figure 1A shows how this is easily accomplished with an 8085 microcomputer system.

Silencing the speech module can be done at any time using the $\overline{MUTE}$ input. The $\overline{MUTE}$ input should be held in an active low state for as long as the silent condition is required.

8085 µp Bus Interface

Figure 1A

Typical I/O Port Interface

Figure 1B

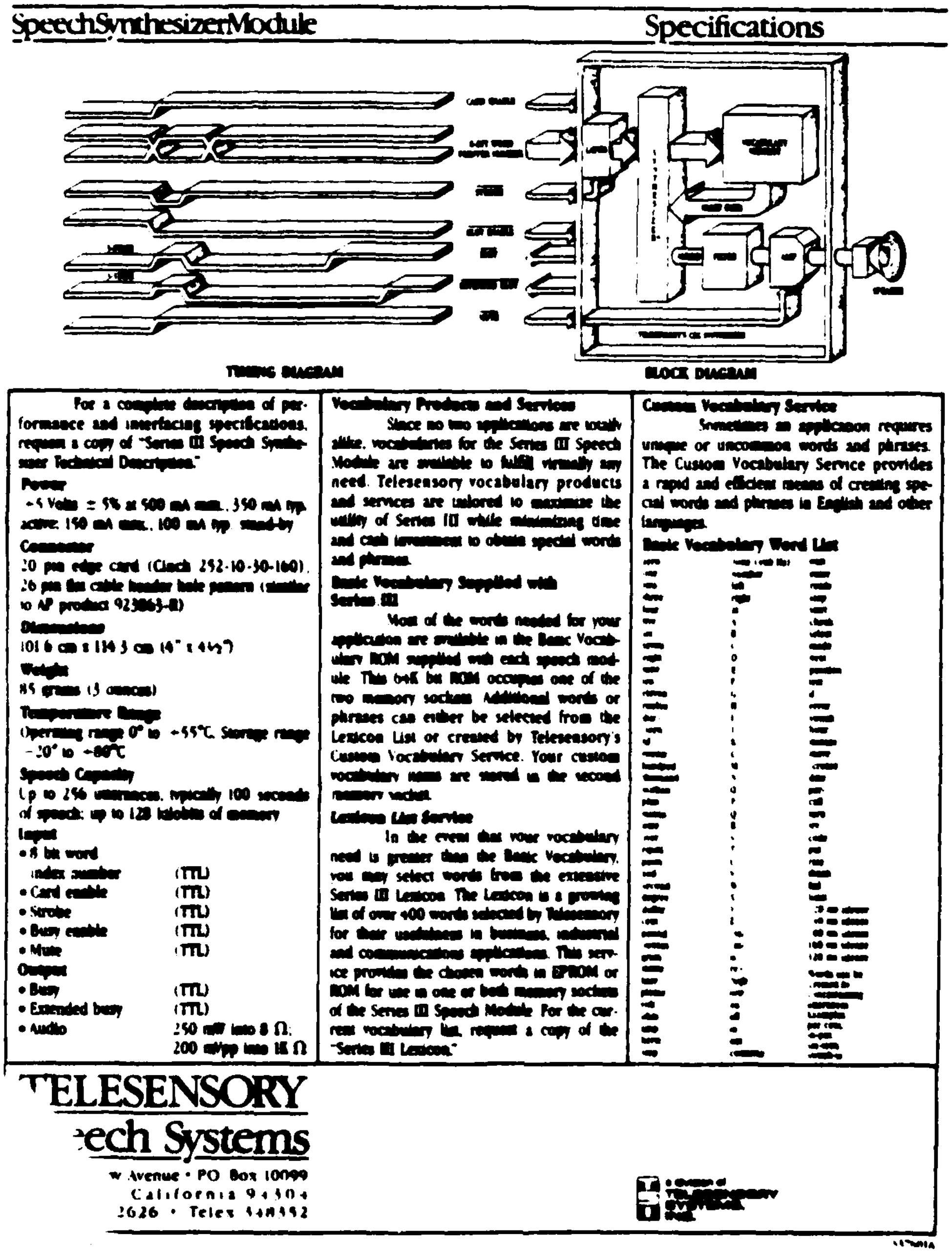
Switch Interface

Figure 1C

- 106 -

TABLE 37

Telesensory Speech Systems Series III Speech Synthesizer Module - Further Specifications



SpeechSynthesizerModule Specifications

TIMING DIAGRAM

BLOCK DIAGRAM

For a complete description of performance and interfacing specifications, request a copy of "Series III Speech Synthesizer Technical Description."

Power

+5 Volts ± 5% at 500 mA max., 350 mA typ. active; 150 mA max., 100 mA typ. stand-by

Connector

20 pin edge card (Cinch 252-10-30-160), 26 pin flat cable header hole pattern (similar to AP product 923863-R)

Dimensions

101.6 cm x 114.3 cm (4" x 4½")

Weight

85 grams (3 ounces)

Temperature Range

Operating range 0° to +55°C. Storage range −20° to +80°C

Speech Capacity

Up to 256 utterances, typically 100 seconds of speech; up to 128 kilobits of memory

Input

• 8 bit word index number	(TTL)
• Card enable	(TTL)
• Strobe	(TTL)
• Busy enable	(TTL)
• Mute	(TTL)

Output

• Busy	(TTL)
• Extended busy	(TTL)
• Audio	250 mW into 8 Ω; 200 mVpp into 1K Ω

Vocabulary Products and Services

Since no two applications are totally alike, vocabularies for the Series III Speech Module are available to fulfill virtually any need. Telesensory vocabulary products and services are tailored to maximize the utility of Series III while minimizing time and cash investment to obtain special words and phrases.

Basic Vocabulary Supplied with Series III

Most of the words needed for your application are available in the Basic Vocabulary ROM supplied with each speech module. This 64K bit ROM occupies one of the two memory sockets. Additional words or phrases can either be selected from the Lexicon List or created by Telesensory's Custom Vocabulary Service. Your custom vocabulary items are stored in the second memory socket.

Lexicon List Service

In the event that your vocabulary need is greater than the Basic Vocabulary, you may select words from the extensive Series III Lexicon. The Lexicon is a growing list of over 400 words selected by Telesensory for their usefulness in business, industrial and communications applications. This service provides the chosen words in EPROM or ROM for use in one or both memory sockets of the Series III Speech Module. For the current vocabulary list, request a copy of the "Series III Lexicon."

Custom Vocabulary Service

Sometimes an application requires unique or uncommon words and phrases. The Custom Vocabulary Service provides a rapid and efficient means of creating special words and phrases in English and other languages.

Basic Vocabulary Word List

[illegible]

TELESENSORY
ech Systems

w Avenue • PO Box 10099
California 94304
2626 • Telex 348352

- 107 -

an additional $30 to $60. Following is Table 38 of the Telesensory mini circuit synthesizer boards.

Telesensory offers a wide variety of LPC synthesizer board-level products. Telesensory particularly emphasizes the wide variety of custom vocabularies that they have available, including numerous foreign languages.

Finally, Telesensory notes that they have several new products which are either available, or coming out soon. First, there is a real-time text-to-speech rule synthesizer. This unit converts ASCII characters to speech, via a cascade/parallel synthesizer, and should cost around $3,500. This unit should be most interesting to evaluate, for it would appear to be the first commercially-available product to incorporate a cascade/parallel synthesizer. Second, Telesensory has just brought out the speech 1020 unit, which is a speech 1000 unit with a self-contained unit with internal power supply. The unit is called the RSC1020, and it sells for approximately $2,500, with vocabulary an additional cost. Telesensory especially emphasizes their custom vocabulary capabilities and their ability to serve customers with relatively low-volume needs.

As mentioned earlier Telesensory Speech Systems has a telephone line for demonstrating their synthesizers: call (415) 969-6257. Their telephone demonstration includes their new text-to-speech unit.

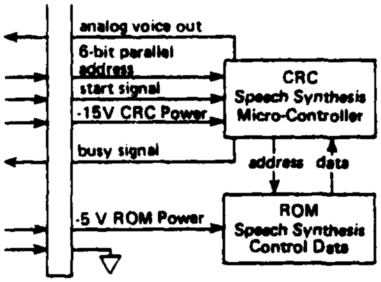
TABLE 38

Telesensory Mini Circuit Synthesizer Boards

Speech Synthesizer Module

DESCRIPTION OF OPERATION

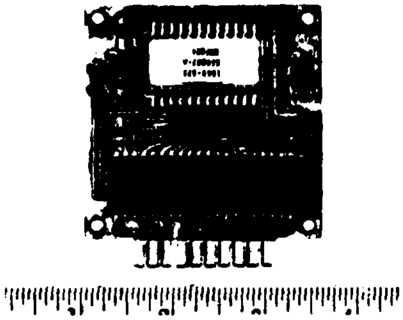
Originally developed for use in TSI's talking calculator for the blind, we are now making our unique speech synthesizer circuit boards available for small computer and OEM applications. Pre-programmed vocabulary data is stored in either one or two 16K MOS ROM.s (depending on the number of words in the vocabulary). When provided with a 6-bit parallel binary address code and a START signal, the custom LSI ROM controller (CRC) fetches appropriate control data from the ROM, determines the speech characteristic of the word, and converts the digital information to an analog audio signal via an on-chip D/A converter. The analog then requires filtering and amplification. The result is a clear, highly intelligible male voice. The operation of the board is described in the block diagram.



A VARIETY OF VOCABULARY CHOICES

Mini Circuit Boards

Mini Circuit Boards are small PC boards measuring less than 3.10" square which provide the minimum necessary components for speech synthesis: the CRC micro-controller, one or two 16K ROM's (depending on the vocabulary selected), and clock frequency circuitry. Vocabularies available include the 24-word calculator vocabularies described under **Calculator Speech Synthesis Module** as well as two 64-word general-purpose vocabularies.



- ***Power:*** -5V and -15V
- In addition to power, an audio filter circuit (described in the Engineering Note which accompanies the board) an audio amplifier, and a speaker must be provided by the user.
- ***Interface:*** Double-sided edge connector, ten pins each side.
- Can be made TTL compatible.

S2A	24-word Calculator Vocabulary	This model also available in French (S2F) and German (S2D)

Calculator Speech Synthesis Module

Features and Specifications

- *Calculator Vocabulary*

oh	four	eight	percent	em (m)	minu
one	five	nine	low	times	plus
two	six	times-minus	over	point	clear
three	seven	equals	root	overflow	swap

Custom Vocabularies

A custom vocabulary can be programmed to fit your particular applications.

Limited Warranty

The Speech Synthesis Module is warranted against defects in material and/or workmanship for a period of 90 days from the date of delivery. Upon specific written request, a copy of the complete product warranty may be obtained free of charge from Telesensory Systems, Inc., at the address stated below

TELESENSORY
Speech Systems
3408 Hillview Avenue • P.O. Box 10099
Palo Alto, California 94304
(415) 493-2626 • Telex: 348352

a division of
TELESENSORY
SYSTEMS,
INC.

S27601D

Texas Instruments

Texas Instruments (TI) markets speech synthesis products which are in the board-level category. The company makes a variety of LPC synthesizers.

Since the synthesizers rely upon an analysis synthesis approach, actual speech signals are analyzed and only the main spectral characteristics are reproduced. TI notes that there are two main types of analysis synthesis synthesizers: 1) formant and 2) LPC. The first synthesizers produced were the basic formant synthesizers, followed by the currently popular LPC synthesizers. For both types of synthesizers, the use of downsampling reduces the bit rate from the original speech, on the order of 100 to 1. This is essential where memory space is limited. However, very low data rates can lead to relatively low quality voice output, so it is important to reproduce the essential acoustic characteristics of human speech. One advantage of LPC synthesizers is that they reduce coarticulation problems associated with rule synthesizers, since they model output based upon real human speech.

Texas Instruments has just introduced three new voice synthesis processors: the TMS5100, the TMS5200, and the TMS5220 chips. Quantity discounts are available for these chips, which otherwise range from approximately $30 to $45.

TI has several voice synthesis memories available for use with their voice synthesis processors. These are the

TMS6100 and TMS6125 chips. These LSI chips are also relatively inexpensive, with quantity discounts available for the O.E.M. market.

TI markets several evaluation kits for their voice output products. First, they have their SPSB1001-011 evaluation board. This is a small board intended for O.E.M. users. The unit, which contains no microprocessor, is capable of synthesizing eight phrases. It uses a 9-volt power supply, and sells for approximately $99. Second, costing an approximate $1,000, is TI's RS232 speech evaluation board. This board is designed to plug into RS232 interfaces and comes with a 25-word vocabulary (expandable to approximately 1000 words with additional ROMs). This board is available only from TI's Regional Technology Centers. Finally, TI markets the S200 series evaluation board for $499. The unit has less memory than the RS232 board, but still has variable intonation.

Texas Instruments also markets microcomputer board products. These include the TM990/306 speech module, with a standard 200-word industrial vocabulary (up to 400 words when mask-programmed ROMs are used for storage). The unit sells for $1,200. It is also available without the standard vocabulary for applications using customer-specified words (as the TM990/306-2). TI notes that this unit will be replaced soon, and that they currently have new voice synthesis products coming out at a rapid rate. A number of these new products will offer additional capabilities, such as allophone dictionaries, variable intonation, sound effects, etc. A further trend in this area will be an increase in performance to price ratio.

Texas Instruments also specializes in custom speech boards, for very specific customer needs. TI stresses its ability to quickly produce custom speech boards (often as rapidly as 9 months). TI markets custom speech boards even for low-volume applications.

TI also notes that they offer courses in speech synthesis. One popular approach has been for customers to purchase an evaluation board, attend TI's course in speech synthesis ($150), and leave this course with a working knowledge of how to get their evaluation board kit operable.

A recent addition to the TI product line has been the talking Loran C Navigator. Loran C is, of course, the U.S. Coast Guard's main navigational system (hyperbolic navigation). The TI9900 and the TI9900N with speech option announce Loran C navigation information in ships and boats. Up to four items may be selected for announcement from the following list:

1) time

2) position

3) speed over the bottom

4) range to waypoint

5) time to go

6) cross-track-error

7) course made good

8) bearing to waypoint

The unit may be set to announce its four messages at intervals ranging from 6 seconds to 1 hour. It announces power-up status, system warnings, and entry corrections. The unit sells for $695 plus installation. This price is very competetive even though the synthesis system is very "special purpose" oriented.

Votrax

Votrax voice synthesis products may be divided into board level and chip level products.

Votrax currently markets at least four products based upon their SC-01CMOS Phoneme Speech Synthesizer. This is essentially a rule synthesizer, which can phonetically synthesize continuous speech, of unlimited vocabulary, from low data rate inputs. This latter point is the main advantage of rule synthesizers, in that such synthesizers typically require large storage areas, which tends to limit the potential size of the output vocabulary. The SC-01 unit consists of a single chip containing 64 different phonemes which are accessed by a 6-bit code. The proper sequential combination of these phoneme codes creates continuous speech. Note that the SC-01 is a very cost-effective unit, priced at $55 (quantities of five or more). Table 39 gives the characteristics of the SC-01 chip.

TABLE 39

Votrax SC-01 Synthesizer Chip

Votrax
A Division of Federal Screw Works
500 Stephenson Highway
Troy, Michigan 48084

SC-01 SPEECH SYNTHESIZER DATA SHEET

Votrax® CMOS Phoneme Speech Synthesizer

GENERAL DESCRIPTION

The SC 01 Speech Synthesizer is a completely self-contained solid state device. This single chip phonetically synthesizes continuous speech of unlimited vocabulary, from low data rate inputs. Figure 1

Speech is synthesized by combining phonemes (the building blocks of speech) in the appropriate sequence. The SC-01 Speech Synthesizer *contains 64 different phonemes* which are accessed by a 6-bit code. It is the proper sequential combination of these phoneme codes that creates continuous speech.

The SC 01 Speech Synthesizer is cost-effective, consumes minimal power and enables in-house product development without vendor dependency. Signals from the SC-01 are applied to an audio output device to amplify and distribute the synthesized speech. See Figure 2.

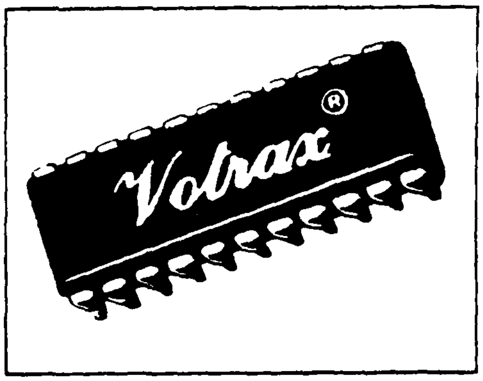


Figure 1 Votrax® SC-01 Speech Synthesizer

FEATURES

- Single CMOS chip
- 70 bits per second
- 22 pin package
- 9 ma. current drain
- Wide voltage supply range
- Latched 5V compatible inputs
- Digital pitch level inputs
- Automatic inflection
- On-chip master clock circuit
- Optional external master clock
- Variety of voice effects
- Sound effects
- Customer product security

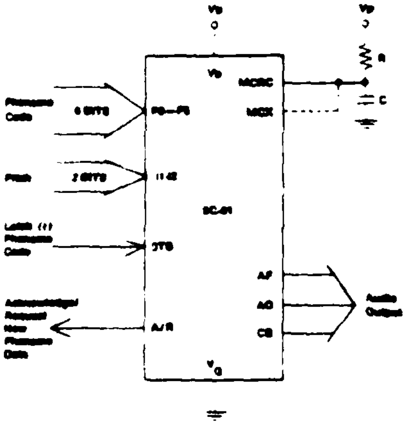


Figure 2 SC-01 Flow Diagram

Votrax also markets the Speech PAC (Phoneme Access Controller) which includes the SC-01 chip. This unit includes provision for additional vocabulary by allowing for storage of additional phoneme codes. Votrax notes that this unit is especially suitable for inclusion with a variety of equipment, controllers, games, etc. The unit contains an EPROM circuit which may be jumpered to accept a 32K EPROM for stored vocabulary expansion. Phonemes and prestored words can be mixed as desired. Following are Tables 40 and 41 documenting features of the Speech PAC unit, which sells for $275, and detailing a flow diagram.

The top-end synthesizer in the Votrax line is the Versatile Speech Module (VSM/1). The unit incorporates additional features over those in the Votrax Speech PAC and sells for $995.

This unit also utilizes the SC-01 synthesizer chip. It has a large lexicon of commonly-used words (industrial engineering based) stored in EPROM. It includes a built-in prefix/suffix table for prestored words. Additional vocabulary can be created and permanently stored on EPROMs (8K to 16K). Other notable features of the unit include a 1,300+ word prestored vocabulary, sound effects, variable stress (4 fixed levels, 4 transitional inflection levels), 8 speech rates, and 8 pause durations. Following are Tables 42 and 43 with a listing of the VSM/1 synthesizer's features, applications, and specifications.

TABLE 40

Votrax Speech Pac

SPEECH PAC T.M.
(Phoneme Access Controller)

Votrax®
PRODUCT DATA

FEATURES

- **Low cost** complete system
- ***Phoneme accessing capability*** for unlimited system vocabulary
- **Additional vocabulary** specific to user needs can be created and permanently stored
- **Ultra low bit rate** of SC-01 maximizes ROM word storage capability
- ***True synthetic speech technology*** eliminates the constraints of a small, fixed vocabulary speech module
- **Parallel interface** for computer, controller or preselected diode matrix to access prestored words or create phonetic speech
- **On board audio amplifier** with volume control

APPLICATIONS

- **Low budget** systems for personal, experimental or low volume OEM product design
- **Fixed vocabulary** for systems requiring limited vocabulary
- **Add on speech output** for existing controllers, educational programs, talking games, etc.
- **Annunciators** for alarm systems, elevators, stations, etc.

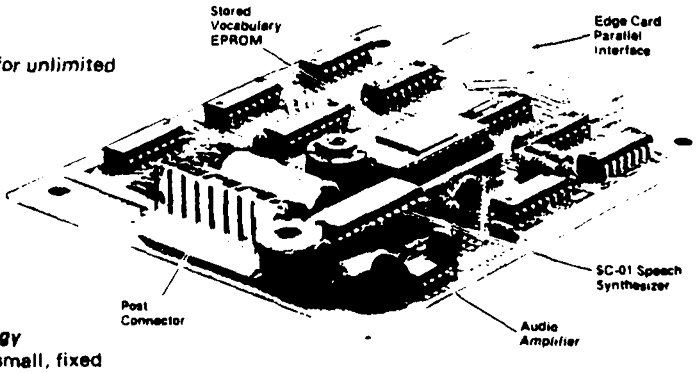


Figure 1. Votrax **Speech PAC**™
(Phoneme Access Controller)

DESCRIPTION

The Votrax **Speech PAC**™ introduces a new level of speech synthesis performance and flexibility at low cost. Based on the truly synthetic speech technology of the SC-01, the **Speech PAC**™ provides the system designer with a small, self-contained circuit board which is easily adapted for use with a variety of equipment, controllers, games, etc.

The **Speech PAC**™ is customer programmable and expandable. The user can easily reconfigure the **Speech PAC**™ vocabulary, as desired. The EPROM socket may be jumpered to accept a 32K EPROM for stored vocabulary expansion. Phonemes and prestored words can be mixed, as desired, to produce an output with unlimited vocabulary.

TABLE 41

Votrax Speech PAC - Further Specifications

SPEECH PAC$^{T.M.}$ – PHONEME ACCESS CONTROLLER

OPERATING CAPABILITY

Prestored words are accessed in 8 byte increments. The low baud rate of the SC-01 Speech Synthesizer allows a single 2716 EPROM to store up to 255 words, and a single 2532 EPROM to store up to 511 words. Long phoneme sequences (more than 8 phonemes) may cross entry boundaries. The **Speech PAC**^{TM} signals the external controller at the end of each phoneme sequence.

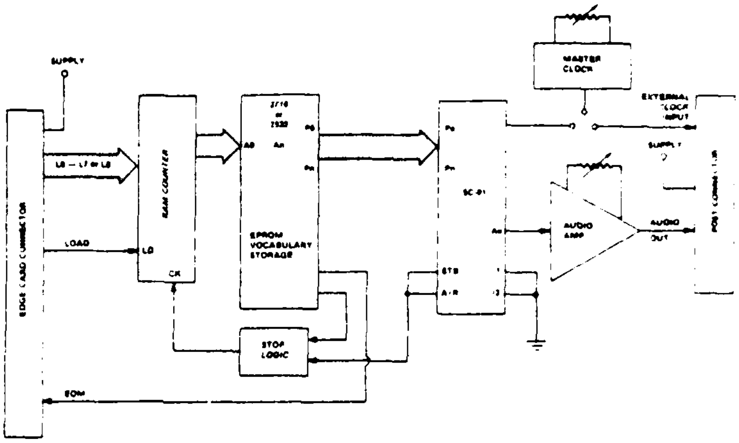


Figure 2. Prestored Word Mode

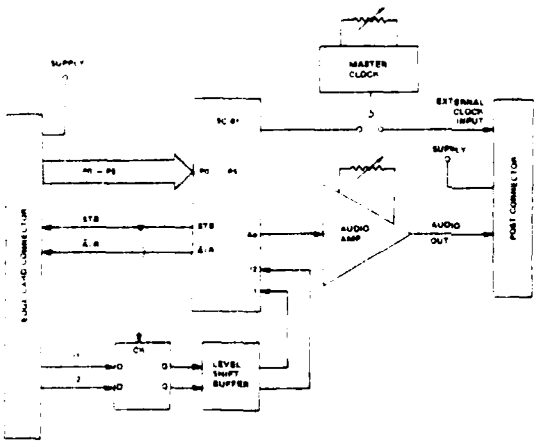


Figure 3. Phoneme Mode

SPECIFICATIONS

- SC-01 Phoneme Synthesizer
- Up to 255 word storage in a single 2716 EPROM
- Expandable with the use of a 32K EPROM
- Mixed prestored word/phrase and phoneme sequencing
- On board audio amplifier
- Parallel interface
- External master clock option
- Handshaking with external controller
- Unlimited vocabulary
- User custom programmable
- Adaptable for use with limit switches and minimal intelligent controllers

TABLE 42

Votrax VSM/1 Synthesizer

VERSATILE SPEECH MODULE T.M (VSM/1)

Votrax®

PRODUCT DATA

FEATURES

- **True synthetic speech technology** and a built in microcomputer eliminate the constraints of a small fixed vocabulary speech module
- **Ultra low bit rate** of the SC-01 maximizes ROM word storage capabilities
- **Large lexicon** of commonly used words with industrial engineering base stored in EPROM
- **Built in prefix/suffix table** for prestored words
- **Additional vocabulary** can be created and permanently stored
- **Phoneme accessing capability** for unlimited vocabulary
- **Speech rate and pitch** dynamic programming for stress patterns and simulation of multi-voice environments
- **Sound effects**, from gunfire to musical sequences can be easily created from prestored sound macros. Additional sound macros can be user defined and EPROM stored for even greater flexibility.
- **Expandable** via interface ports
- **Parallel and RS232 compatible serial interfacing** with selectable baud rates and terminal modes
- **Foreground and background** simultaneous operation for speech and voxOS (voice operating system)
- **Built in microcomputer** can also simultaneously perform monitoring activities and execute speech commands

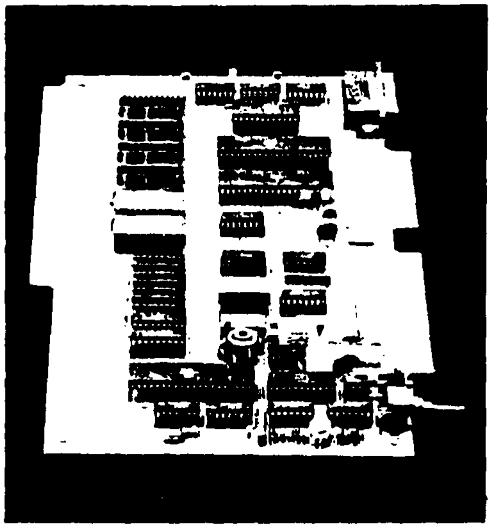


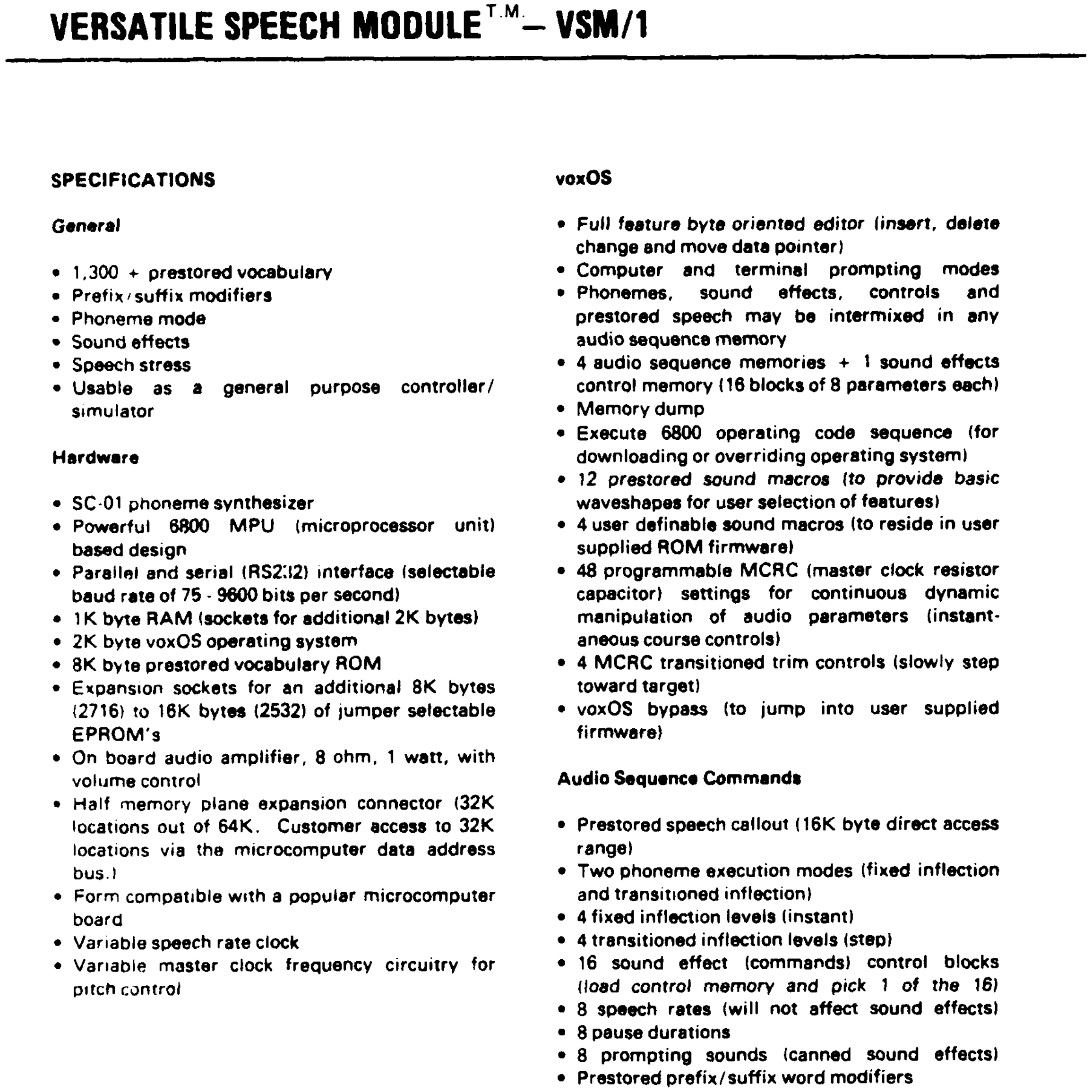
Figure 1. Votrax **VSM/1™**
(Versatile Speech Module)

APPLICATIONS

- The **VSM/1™** can be used as a **microcomputer** to simulate or develop talking products, such as a talking calculator or talking games. It can also be used for unlimited real time speech synthesis while simultaneously executing commands and performing monitoring activities.
- The **VSM/1™** can plug directly into the card cage of an industrial control computer to provide **prompting for operating personnel** (instructions for a real time situation). Typical applications are chemical processing plants, nuclear power stations, aircraft systems, seismic monitoring stations and automated warehousing.

TABLE 4.3

Votrax VSM/1 Synthesizer -- Further Specifications



VERSATILE SPEECH MODULE T.M. – VSM/1

SPECIFICATIONS

General

- 1,300 + prestored vocabulary
- Prefix/suffix modifiers
- Phoneme mode
- Sound effects
- Speech stress
- Usable as a general purpose controller/simulator

Hardware

- SC-01 phoneme synthesizer
- Powerful 6800 MPU (microprocessor unit) based design
- Parallel and serial (RS232) interface (selectable baud rate of 75 - 9600 bits per second)
- 1K byte RAM (sockets for additional 2K bytes)
- 2K byte voxOS operating system
- 8K byte prestored vocabulary ROM
- Expansion sockets for an additional 8K bytes (2716) to 16K bytes (2532) of jumper selectable EPROM's
- On board audio amplifier, 8 ohm, 1 watt, with volume control
- Half memory plane expansion connector (32K locations out of 64K. Customer access to 32K locations via the microcomputer data address bus.)
- Form compatible with a popular microcomputer board
- Variable speech rate clock
- Variable master clock frequency circuitry for pitch control

voxOS

- *Full feature byte oriented editor (insert, delete change and move data pointer)*
- Computer and terminal prompting modes
- Phonemes, sound effects, controls and prestored speech may be intermixed in any audio sequence memory
- 4 audio sequence memories + 1 sound effects control memory (16 blocks of 8 parameters each)
- Memory dump
- Execute 6800 operating code sequence (for downloading or overriding operating system)
- *12 prestored sound macros (to provide basic* waveshapes for user selection of features)
- 4 user definable sound macros (to reside in user supplied ROM firmware)
- 48 programmable MCRC (master clock resistor capacitor) settings for continuous dynamic manipulation of audio parameters (instantaneous course controls)
- 4 MCRC transitioned trim controls (slowly step toward target)
- voxOS bypass (to jump into user supplied firmware)

Audio Sequence Commands

- Prestored speech callout (16K byte direct access range)
- Two phoneme execution modes (fixed inflection and transitioned inflection)
- 4 fixed inflection levels (instant)
- 4 transitioned inflection levels (step)
- 16 sound effect (commands) control blocks *(load control memory and pick 1 of the 16)*
- 8 speech rates (will not affect sound effects)
- 8 pause durations
- 8 prompting sounds (canned sound effects)
- Prestored prefix/suffix word modifiers

The final Votrax product that SCRL reviewed is the Type-'N-Talk text-to-speech synthesizer. The unit costs $375. To use this system, words are typed into a host terminal and translated into synthesized speech by the system's microprocessor-based text-to-speech algorithm. The unit incorporates the SC-01 chip. It includes a 1-watt amplifier, RS 232C interface, data echo of ASCII characters, and phoneme access modes. Table 44 provides a basic description of the Type-'N-Talk unit.

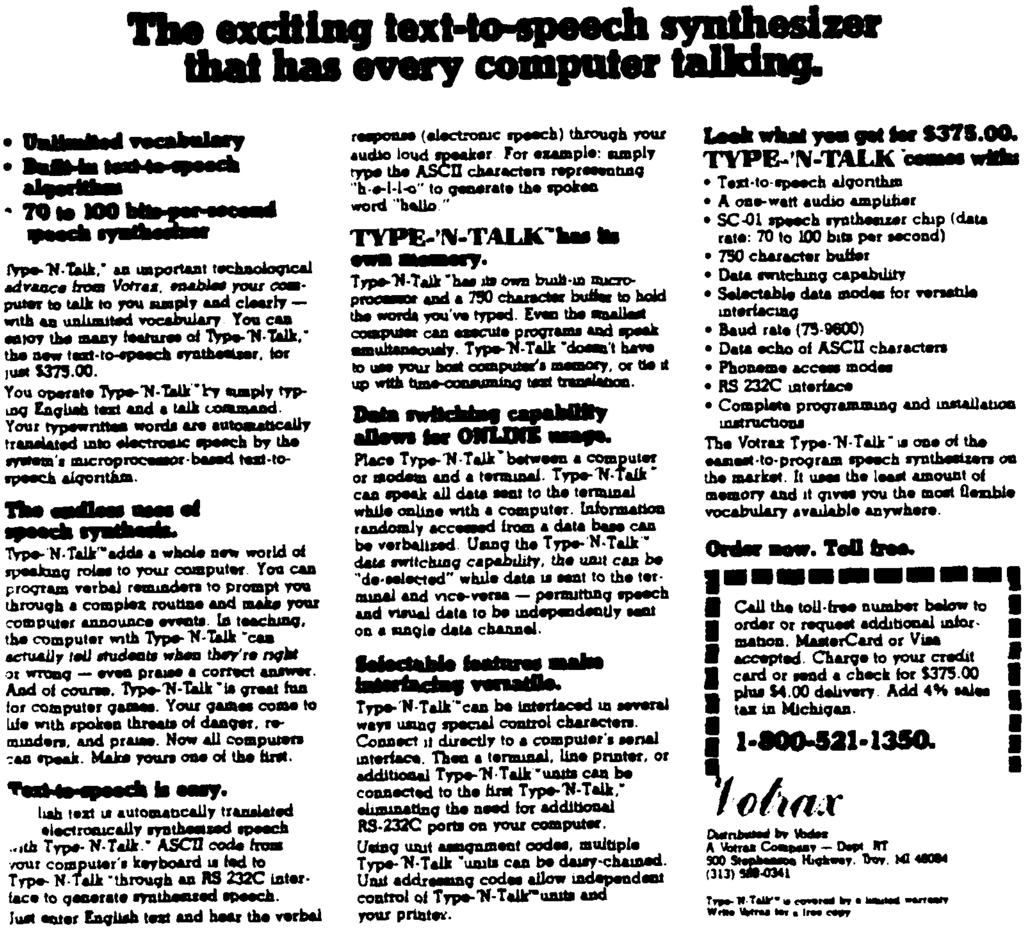
Note that Votrax has just announced a second text-to-speech synthesizer, with reportedly better voice quality than their Type-'N-Talk unit. This is the SVA text-to-speech unit, which sells for approximately $1,650. The unit is available with a 16K buffer (approximately 15 characters per second), which will hold up to 800 characters. This Votrax unit is also a rule synthesizer.

5.2 KLATT SYNTHESIZER PROGRAM

This subsection deals with the Klatt synthesizer program. This is not yet a marketed synthsizer as were the ones discussed in the proceeding section. The March, 1980, Journal of the Acoustical Society of America contained an article on Dr. Klatt's parametric (rule) synthesizer. This synthesizer is the most advanced rule synthesizer to date, and output from Klatt's synthesizer is virtually indistinguishable from live speech, given the appropriate parameter input.

TABLE 44

Votrax Type-'N-Talk Text to Speech Synthesizer



The exciting text-to-speech synthesizer that has every computer talking.

- **Unlimited vocabulary**
- **Built-in text-to-speech algorithm**
- **70 to 100 bits-per-second speech synthesizer**

Type-'N-Talk,™ an important technological *advance from Votrax, enables your computer* to talk to you simply and clearly — with an unlimited vocabulary. You can enjoy the many features of Type-'N-Talk,™ the new text-to-speech synthesizer, for just $375.00.

You operate Type-'N-Talk™ by simply typing English text and a talk command. Your typewritten words are automatically translated into electronic speech by the system's microprocessor-based text-to-speech algorithm.

The endless uses of speech synthesis.

Type-'N-Talk™ adds a whole new world of speaking roles to your computer. You can program verbal reminders to prompt you through a complex routine and make your computer announce events. In teaching, the computer with Type-'N-Talk™ can *actually tell students when they're right* or wrong — even praise a correct answer. And of course, Type-'N-Talk™ is great fun for computer games. Your games come to life with spoken threats of danger, reminders, and praise. Now all computers can speak. Make yours one of the first.

Text-to-speech is easy.

lish text is automatically translated electronically synthesized speech with Type-'N-Talk.™ ASCII code from your computer's keyboard is fed to Type-'N-Talk™ through an RS 232C interface to generate synthesized speech. Just enter English text and hear the verbal response (electronic speech) through your audio loud speaker. For example: simply type the ASCII characters representing "h-e-l-l-o" to generate the spoken word "hello."

TYPE-'N-TALK™ has its own memory.

Type-'N-Talk™ has its own built-in microprocessor and a 750 character buffer to hold the words you've typed. Even the smallest computer can execute programs and speak simultaneously. Type-'N-Talk™ doesn't have to use your host computer's memory, or tie it up with time-consuming text translation.

Data switching capability allows for ONLINE usage.

Place Type-'N-Talk™ between a computer or modem and a terminal. Type-'N-Talk™ can speak all data sent to the terminal while online with a computer. Information randomly accessed from a data base can be verbalized. Using the Type-'N-Talk™ *data switching capability, the unit can be* "de-selected" while data is sent to the terminal and vice-versa — permitting speech and visual data to be independently sent on a single data channel.

Selectable features make interfacing versatile.

Type-'N-Talk™ can be interfaced in several ways using special control characters. Connect it directly to a computer's serial interface. Then a terminal, line printer, or additional Type-'N-Talk™ units can be connected to the first Type-'N-Talk,™ eliminating the need for additional RS-232C ports on your computer.

Using unit assignment codes, multiple Type-'N-Talk™ units can be daisy-chained. Unit addressing codes allow independent control of Type-'N-Talk™ units and your printer.

Look what you get for $375.00. TYPE-'N-TALK™ comes with:

- Text-to-speech algorithm
- A one-watt audio amplifier
- SC-01 speech synthesizer chip (data rate: 70 to 100 bits per second)
- 750 character buffer
- Data switching capability
- Selectable data modes for versatile interfacing
- Baud rate (75-9600)
- Data echo of ASCII characters
- Phoneme access modes
- RS 232C interface
- Complete programming and installation instructions

The Votrax Type-'N-Talk™ is one of the easiest-to-program speech synthesizers on the market. It uses the least amount of memory and it gives you the most flexible vocabulary available anywhere.

Order now. Toll free.

Call the toll-free number below to order or request additional information. MasterCard or Visa accepted. Charge to your credit card or send a check for $375.00 plus $4.00 delivery. Add 4% sales tax in Michigan.

1-800-521-1350.

Votrax

Distributed by Vodex
A Votrax Company — Dept. NT
500 Stephenson Highway, Troy, MI 48084
(313) 588-0341

Type-'N-Talk™ is covered by a limited warranty
Write Votrax for a free copy

Klatt's synthesizer provides minute control over the main parameters underlying human speech. Consequently, anything can potentially be synthesized, given the correct parameter input. This type of synthesizer will certainly see wide commercial application in the future. Figure 4 gives a flow diagram of the Klatt synthesizer.

The synthesizer is a cascade/parallel formant synthesizer as shown in the top schematic of Figure 5. The two main components of the synthesizer are the cascade portion and the parallel portion; this amounts to a combination of the two common types of experimental synthesizers widely seen in the literature.

Parallel synthesizers which are essentially formant resonators that simulate the transfer function of the vocal tract, connected in parallel, are of the type shown in the bottom schematic of Figure 5. Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a formant in the output spectrum for both voiced and voiceless speech sounds. The cascade configuration is noted by Dr. Klatt to have the advantage of having the relative amplitudes of formants automatically computed without the need for individual amplitude controls for each formant. The disadvantage of cascade synthesizers is that one still needs a parallel formant configuration for the generation of fricatives and plosives. This is due to the fact that the vocal tract transfer function cannot be

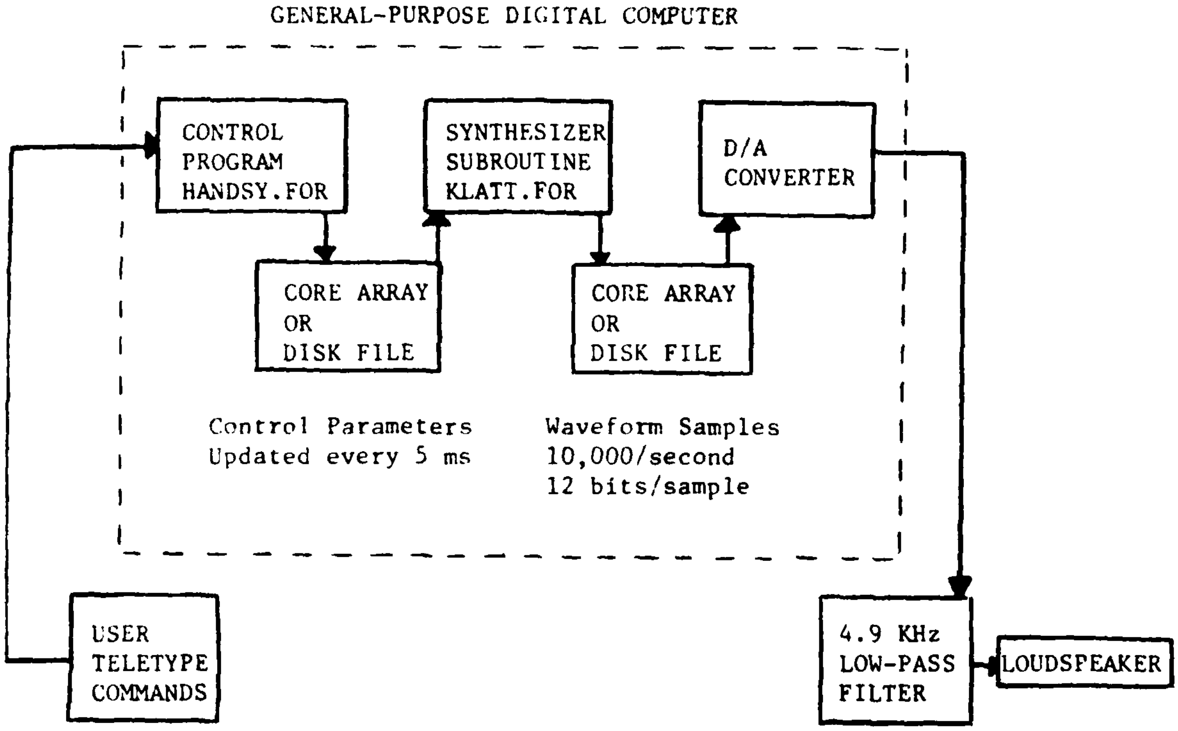
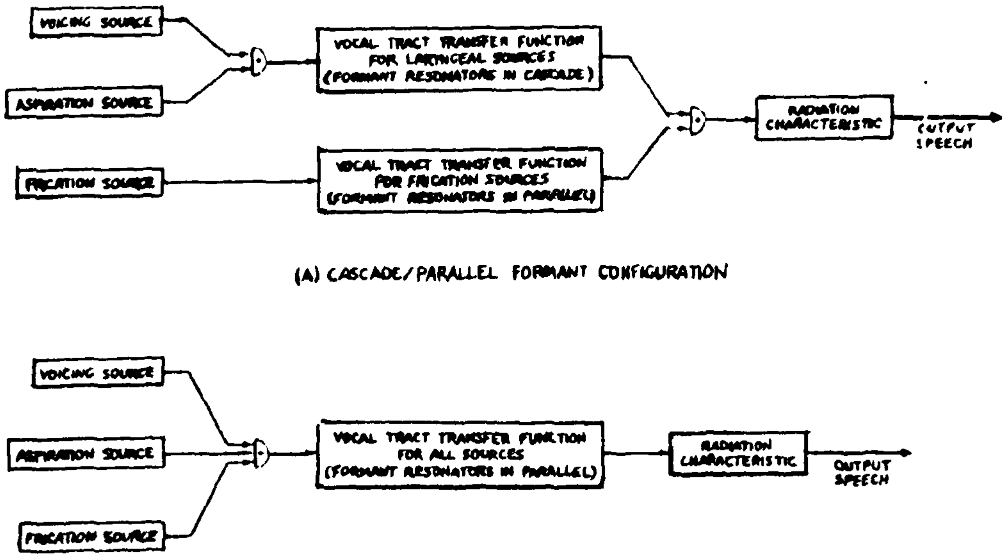

GENERAL-PURPOSE DIGITAL COMPUTER
CONTROL PROGRAM HANDSY.FOR
SYNTHESIZER SUBROUTINE KLATT.FOR
D/A CONVERTER
CORE ARRAY OR DISK FILE
CORE ARRAY OR DISK FILE
Control Parameters Updated every 5 ms
Waveform Samples 10,000/second 12 bits/sample
USER TELETYPE COMMANDS
4.9 KHz LOW-PASS FILTER
LOUDSPEAKER

Figure 4: Flow diagram of Klatt's synthesizer.


VOICING SOURCE
ASPIRATION SOURCE
FRICATION SOURCE
VOCAL TRACT TRANSFER FUNCTION FOR LARYNGEAL SOURCES (FORMANT RESONATORS IN CASCADE)
VOCAL TRACT TRANSFER FUNCTION FOR FRICATION SOURCES (FORMANT RESONATORS IN PARALLEL)
RADIATION CHARACTERISTIC
OUTPUT SPEECH
VOICING SOURCE
ASPIRATION SOURCE
FRICATION SOURCE
VOCAL TRACT TRANSFER FUNCTION FOR ALL SOURCES (FORMANT RESONATORS IN PARALLEL)
RADIATION CHARACTERISTIC
OUTPUT SPEECH

(A) CASCADE/PARALLEL FORMANT CONFIGURATION

(B) SPECIAL-PURPOSE ALL-PARALLEL FORMANT CONFIGURATION

* The synthesizer is normally used in a cascade/parallel configuration shown at the top, but may be used in an all parallel version shown at the bottom if one wishes to exercise independent control over formant amplitudes for vowels.

Figure 5: Configurations for cascade/parallel formant synthesizers and for special-purpose all-parallel formant synthesizers.

modeled adequately by five cascade resonators when the source sound is above the larynx. So, overall cascade synthesizers tend to be relatively more complex. A second advantage of the cascade configuration is that it is a more accurate model of the vocal tract transfer function during the production of non-nasal voiced sounds. Also, it is difficult to match the transfer function of certain vowels using a parallel formant synthesizer.

Klatt's synthesizer uses two voicing sources, one for periodic sounds and one for nonperiodic or turbulent sounds (such as for fricatives). The Klatt synthesizer has a sampling rate of 10000 bps, as speech does not have much energy above 5000 Hz. and low-pass filtered speech sounds perfectly natural.

Klatt's synthesizer has a set of 39 control parameters which are used for synthesis; as many as 20 of these parameters may be used for English utterances. The Klatt synthesizer basically uses the parameters for input, functioning as a digital resonator. Tables 45 and 46 list variable parameters and sample parameters.

Spectrograms are used by Klatt as a model to determine the general acoustic characteristics of the utterance to be synthesized. Spectrograms of natural speech, compared with synthesized speech, show just how well the Klatt synthesizer models human speech. Figure 6 displays a comparison between a natural utterance and a synthesized utterance.

TABLE 45

Klatt Synthesizer Variable Parameters

TABLE List of control parameters for the software formant synthesizer. The second column indicates whether the parameter is normally constant (C) or variable (V) during the synthesis of English sentences. Also listed are the permitted range of values for each parameter, and a typical constant value.

N	V/C	Sym	Name	Min	Max	Typ
1	V	AV	Amplitude of voicing (dB)	0	80	0
2	V	AF	Amplitude of frication (dB)	0	80	0
3	V	AH	Amplitude of aspiration (dB)	0	80	0
4	V	AVS	Amplitude of sinusoidal voicing (dB)	0	80	0
5	V	F0	Fundamental freq. of voicing (Hz)	0	500	0
6	V	F1	First formant frequency (Hz)	150	900	450
7	V	F2	Second formant frequency (Hz)	500	2500	1450
8	V	F3	Third formant frequency (Hz)	1300	3500	2450
9	V	F4	Fourth formant frequency (Hz)	2500	4500	3300
10	V	FNZ	Nasal zero frequency (Hz)	200	700	250
11	C	AN	Nasal formant amplitude (dB)	0	80	0
12	C	A1	First formant amplitude (dB)	0	80	0
13	V	A2	Second formant amplitude (dB)	0	80	0
14	V	A3	Third formant amplitude (dB)	0	80	0
15	V	A4	Fourth fromant amplitude (dB)	0	80	0
16	V	A5	Fifth formant amplitude (dB)	0	80	0
17	V	A6	Sixth formant amplitude (dB)	0	80	0
18	V	AB	Bypass path amplitude (dB)	0	80	0
19	V	B1	First formant bandwidth (Hz)	40	500	50
20	V	B2	Second formant bandwidth (Hz)	40	500	70
21	V	B3	Third formant bandwidth (Hz)	40	500	110
22	C	SW	Cascade/parallel switch	0(CASC)	1(PARA)	0
23	C	FGP	Glottal resonator 1 frequency (Hz)	0	600	0
24	C	BGP	Glottal resonator 1 bandwidth	100	2000	100
25	C	FGZ	Glottal zero frequency (Hz)	0	5000	1500
26	C	BGZ	Glottal zero bandwidth (Hz)	100	9000	6000
27	C	B4	Fourth formant bandwidth (Hz)	100	500	250
28	V	F5	Fifth formant frequency (Hz)	3500	4900	3750
29	C	B5	Fifth formant bandwidth (Hz)	150	700	200
30	C	F6	Sixth formant frequency (Hz)	4000	4999	4900
31	C	B6	Sixth formant bandwidth (Hz)	200	2000	1000
32	C	FNP	Nasal pole frequency (Hz)	200	500	250
33	C	BNP	Nasal pole bandwidth (Hz)	50	500	100
34	C	BNZ	Nasal zero bandwidth (Hz)	50	500	100
35	C	BGS	Glottal resonator 2 Bandwidth	100	1000	200
36	C	SR	Sampling rate	5000	20000	10000
37	C	NWS	No. of waveform samples per chunk	1	200	50
38	C	GO	Overall gain control (dB)	0	800	47
39	C	NFC	Number of cascaded formants	4	6	5

TABLE 46

Klatt Synthesizer Sample Parameters

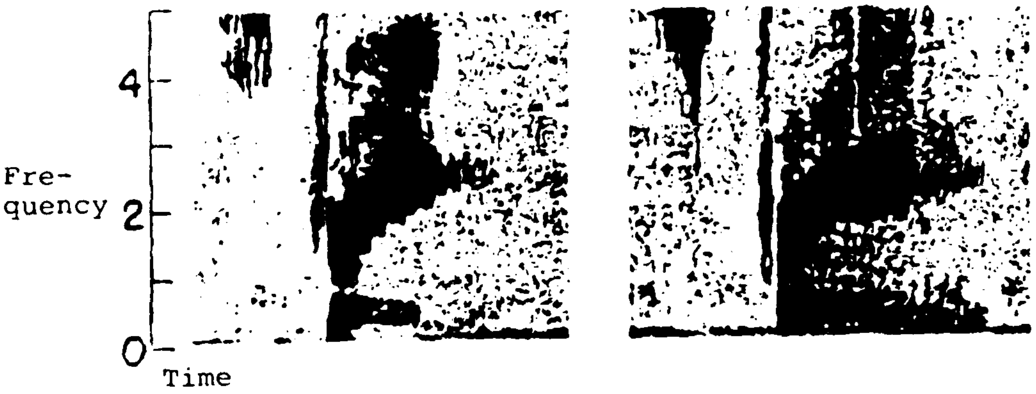
KLATT CASCADE/PARALLEL FORMAT SYNTHESIZER

THE FOLLING TABLE REPRESENTS THE CONFIGURATION FOR THE CURRENT PARAMETER FILE.

NUM	PARM	V/C	VALUE	NUM	PARM	V/C	VALUE	NUM	PARM	V/C	VALUE
1	AV	1	0	14	A3	1	60	27	B4	0	3400
2	AF	1	0	15	A4	1	60	28	F5	0	3700
3	AH	1	0	16	A5	1	0	29	B5	0	500
4	AVS	1	0	17	A6	1	0	30	F6	0	4900
5	F0	1	0	18	AB	1	0	31	B6	0	800
6	F1	1	450	19	B1	1	50	32	FNP	0	250
7	F2	1	1450	20	B2	1	70	33	BNP	0	100
8	F3	1	2450	21	B3	1	110	34	BNZ	0	100
9	F4	1	3300	22	SW	0	0	35	FRA	0	99
10	FNZ	1	250	23	FGP	0	0	36	SR	0	10000
11	AN	1	0	24	BGP	0	150	37	NWS	0	50
12	A1	1	60	25	FGZ	0	1500	38	G0	0	66
13	AZ	1	60	26	BGZ	0	6000	39	NF	0	5

NUM	AV	AF	AH	AVS	F0	F1	F2	F3	F4	FNZ	AN	A1	A2
0	20	0	0	0	100	450	1450	2450	3300	250	0	60	60
5	27	0	0	0	105	450	1450	2450	3300	250	0	60	60
10	33	0	0	0	110	450	1450	2450	3300	250	0	60	60
15	41	0	0	0	115	450	1450	2450	3300	250	0	60	60
20	50	0	0	0	120	450	1450	2450	3300	250	0	60	60
25	51	0	0	0	125	450	1450	2450	3300	250	0	60	60
30	52	0	0	0	130	450	1450	2450	3300	250	0	60	60
35	54	0	0	0	135	450	1450	2450	3300	250	0	60	60
40	55	0	0	0	140	450	1450	2450	3300	250	0	60	60
45	56	0	0	0	145	450	1450	2450	3300	250	0	60	60
50	58	0	0	0	150	450	1450	2450	3300	250	0	60	60
55	59	0	0	0	145	450	1450	2450	3300	250	0	60	60
60	61	0	0	0	139	450	1450	2450	3300	250	0	60	60
65	62	0	0	0	134	450	1450	2450	3300	250	0	60	60
70	63	0	0	0	128	450	1450	2450	3300	250	0	60	60
75	65	0	0	0	123	450	1450	2450	3300	250	0	60	60
80	50	0	0	0	117	450	1450	2450	3300	250	0	60	60
85	50	0	0	0	112	450	1450	2450	3300	250	0	60	60
90	43	0	0	0	106	450	1450	2450	3300	250	0	60	60
95	35	0	0	0	100	450	1450	2450	3300	250	0	60	60

A3	A4	A5	A6	AB	B1	B2	B3
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110
60	60	0	0	0	50	70	110



Broadband spectrograms are compared of a natural and synthetic word, "string," spoken by a female talker.

Figure 6: Natural utterance compared to Klatt's synthesized utterance.

Klatt states the usefulness of the linear prediction spectrum. To obtain this spectrum, a linear prediction analysis precedes a discrete Fourier transform. The autocorrelation alogrithm (Makhoul, 1975) using 14 poles, is applied.

It is expected that Klatt's current rule synthesizer will be fully incorporated into future commercial products, with appropriate control software. A synthesizer incorporating Klatt's latest synthesis would probably include storage capabilities for control parameters. Finally, it is expected that once Klatt's current synthesis program is incorporated into a commercial synthesizer, it should provide very serious competition for currently available synthesizers, because the output is often not easily distinguishable from live speech.

CHAPTER 6.

CONCLUSIONS FOR USE OF SPEECH RECOGNITION AND SPEECH SYNTHESIS TECHNIQUES

This chapter examines and compares different speech recognition and speech synthesis technologies as they relate to Coast Guard operational and technical requirements. It also discusses problems involved in developmental efforts for speech recognition and synthesis. Finally, it proposes a future plan for using speech synthesis for broadcasting Coast Guard weather reports.

6.1 SPEECH RECOGNITION TECHNOLOGY

Coast Guard operational and technical requirements lead us to the conclusion that the Coast Guard requires a totally speaker-independent speech recognition system capable of spotting keywords in connected speech. In particular, the Coast Guard has considered speech recognition as a potential means to back up watch standers in guarding distress frequencies. Application areas mentioned by the Coast Guard included: Communications Stations, Radio Stations, Group Stations, Search and Rescue Stations, and Coast Guard Cutters. Coverage is for 2182 kHz MF radiotelephone, and 156.8 mHz radiotelephone (Channel 16). We pointed out that our analysis of selected Coast Guard radio transmissions showed that the Coast Guard needs a recognition system capable of handling transmissions with a relatively low signal-to-noise ratio (mean S/N ratio was 23dB, with a standard deviation of 5 dB.), and with cutoff frequencies ranging from approximately 300-4000 Hz. Additional technical requirements make its speech recognition requirements even more stringent. For example, we noted that Coast Guard requirements regarding keyword spotting indicate that a recognizer should be able to handle connected speech input with widely differing emotional states, diverse accents, and substantial nonperiodic background noise input.

As noted previously, SCRL was able to compile detailed information regarding speech recognition products from nine major manufacturers:

1) Centigram
2) Heuristics
3) Interstate Electronics
4) Nippon Electric Company
5) Scott Instruments
6) Threshold Technology
7) Verbex
8) Voicetek
9) Votan

As mentioned earlier, there are no currently available recognizers which can handle connected speech in a completely speaker-independent manner. We did point out that three manufacturers market recognizers capable of handling connected speech. However, of these three manufacturers only one markets a speaker-independent recognizer (capable of recognizing digits plus 50 optional words). Yet this particular recognizer will not handle connected speech input. From our discussions with manufacturers and review of ongoing work related to speech recognition technology, we note that there is a large effort being devoted to the task of developing speaker-independent recognition systems capable of handling connected speech input.

In line with a general price reduction in speech recognition systems due to improved technology and manufacturing techniques, we believe that the price of future speaker-independent recognizers capable of handling continuous speech input will be lower than might first be imagined, probably on the order of $20K, depending upcn the size of the recognition vocabulary. It can also be pointed out that first-generation speaker-independent, continuous speech input voice recognizers will handle the digits primarily, plus several control words. This configuration appears to have wide marketing possibilities related to business usage over the telephone.

It is interesting that three recognizers which would be closest to meeting Coast Guard requirements involve a generally similar approach to voice recognition. Each unit digitally encodes voice input samples for comparison with stored "templates"; the approach is to correlate stored filter coefficients or other stored "template" data with incoming speech samples. Given an appropriately high correlation, a match occurs, the word is recognized, and appropriate ASCII messages are output from the recognition unit. This ASCII output can be used to define a variety of instructions as required. Only one of the recognizers does not require "training", where specific speakers follow a predefined sequence for encoding their recognition vocabulary "templates."

In terms of meeting specific Coast Guard operational and technical requirements regarding spotting of keywords in incoming distress signals ("mayday", "sinking", etc.), we note that the Verbex 1800 comes closest. Again, we note that at least one device might be close. This recognizer can handle up to 50 words plus the digits, in a speaker-independent mode over the telephone (and with accompanying adverse noise conditions) with high accuracy, but it cannot handle connected speech input. It should be pointed out, however, that there may be a possibility of using a speaker-independent, isolated word recognizer for meeting Coast Guard speech recognition requirements.

Although all sample radio transmissions which were acoustically evaluated by SCRL involved connected speech, we do have the impression that mariners generally pronounce distress signals slowly, and repeat keywords such as "mayday". If this is generally the case with incoming Coast Guard distress calls, an isolated word recognizer could be expected to successfully recognize a high percentage of keywords in Coast Guard distress signals. We should point out that there remain several uncertainties regarding such an approach. One important consideration would be how false recognitions might be generated by connected speech input surrounding the "isolated" keywords to be spotted.

There are at least two further points to note regarding Coast Guard station automation plans and the possibility of using speech recognition to spot keywords in incoming distress signals. First, there is the basic consideration as to what level of speech recognition product might most easily be integrated into Coast Guard station automation plans. Second, is the consideration as to the relative cost of different levels of speech recognition devices.

As noted previously,there are several levels of speech recognition products widely seen in today's commercial market, including board-level recognition products and stand-alone devices. Board-level recognition products are most suitable for installations which are already equipped with a host computer capable of accepting the common RS232C interface. Stand-alone recognizers are most suited to installations which do not have a host computer with enough storage to handle speech recognition.

It should be pointed out that stand-alone speech recognition units tend to be relatively expensive, since they are generally built around an existing minicomputer. Wherever this minicomputer can be used for tasks in addition to speech recognition (such as word processing, data storage, etc.), it makes a more cost-efficient package than if it operates only as a speech recognition host.

6.2 SPEECH SYNTHESIS TECHNOLOGY

As pointed out previously, SCRL is very optimistic that presently available speech synthesis technology exists which is fully capable of meeting Coast Guard operational and technical requirements for synthesis of weather reports, notices to mariners, hydrographic information, and other desired broadcasts. Application areas include Communication Stations, Radio Stations, Group Stations, and Search and Rescue Stations. We recommend that the Coast Guard consider using both an analysis synthesizer to obtain natural sounding speech and also an ASCII-prompted (for incoming weather reports via teletype) rule-type "text-to-speech" speech synthesizer for future efficiency and extended ability.

It should be pointed out that the Coast Guard requires a speech synthesizer with an essentially unlimited vocabulary (for names of ships, storms, etc.). Coast Guard speech synthesis also needs high-quality audio output. Synthesized Coast Guard broadcasts should have good prosodics, such as realistic sentence intonation, and a variety of voices. A Coast Guard consideration has been that advanced technologies, such as speech synthesis, offer the possibility of conserving manpower, and thus saving operating funds.

6.3 DEVELOPMENTAL EFFORTS FOR SPEECH RECOGNITION TECHNOLOGY

As this report has mentioned, there are no currently available speech recognizers which can handle connected speech in a speaker-independent mode. A variety of manufacturers are now developing this type of speech recognition system, for it has such a diversity of potential markets.

Manufacturers point out that the development of a completely speaker-independent recognizer that would handle connected speech involves several technical problems which have not yet been fully solved. These problems include:

1) Need for segmentation programs which correctly determine the words or phrases to be matched with stored templates.

2) Need for better time warping alogrithms to more accurately match reference templates with input data.

3) Need for better alogrithms for relating essentially unique or idiosyncratic acoustic manifestations to common reference templates, including both phonetic and prosodic phenomena.

In terms of Coast Guard developmental efforts in the area of speech recognition, this report suggests that the above problems are exceedingly difficult and commercial manufacturers are now working on them. We feel that developmental efforts the Coast Guard might make in this area would closely parallel those of commercial manufacturers of recognizers and would not be cost effective.

Speaker-independent recognition with connected speech should not be that far distant by current manufacturers. We have already pointed out, for example, that one recognizer will handle isolated digits and control words, plus up to 50 selected optional vocabulary items, in a completely speaker-independent mode over the telephone. Several manufacturers of recognizers already market recognizers which will handle connected speech (approximately 180 words per minute). Speech recognition technology is very dependent on the LSI chip industry. As prices for LSI chips continue to decrease, we can expect to see improved performance from commercial speech recognition devices. With this in mind, one can readily envision a completely speaker-independent recognizer which will handle connected speech in the not-too-distant future. For this reason, we suggest that the Coast Guard need not sponsor the development of a special purpose speaker-independent connected speech recognizer for their applications.

One area where the Coast Guard should concentrate its efforts concerns the overall planning strategy for station automation requirements. Several levels of speech recognition products currently exist, for example, board products and complete stand-alone products. To easily integrate speech recognition technology in automation planning, the Coast Guard should consider what kind of an approach it will be taking with regard to computer selection and implementation. A key point to consider concerns the question of how much computing capability is required to meet Coast Guard requirements for automation of its facilities.

6.4 DEVELOPMENTAL EFFORTS FOR SPEECH SYNTHESIS TECHNOLOGY

As we have already stated, speech synthesis technology currently exists which appears capable of meeting Coast Guard operational and technical requirements related to its broadcast requirements. Both an analysis synthesizer and a rule-based text-to-speech synthesizer would provide a convenient means of preparing Coast Guard weather reports, hydrographic information, notices to mariners, safety messages, and other desired broadcasts. Consequently, there is no real need for the Coast Guard to initiate developmental efforts regarding speech synthesis systems for meeting its broadcast requirements. Applications studies are what would be recommended.

As with speech recognition technology, speech synthesis technology is heavily tied to the economics of the LSI chip industry. Recent years have seen a real decline in prices for LSI chips and an increase in their capabilities. This trend is expected to continue, so that speech synthesis technology will become even more attractive, not only in terms of performance, but in terms of price as well.

We do suggest that the Coast Guard consider speech synthesis technology in the framework of its overall automation planning requirements, so that this technology can easily integrate with the overall Coast Guard computing requirements for station automation planning.

6.5 SPEECH RECOGNITION COST EFFECTIVENESS

Speech recognition technology has generally been most cost effective: 1) in terms of the manner in which it increases the efficiency of human/equipment or man/machine interactions, 2) through its ability to automate procedures that previously required human operators, and 3) in terms of its overall efficiency in recording data from human operators.

1) Speech recognition technology has been cost effective for humans who are employed in situations where their hands are occupied, but they must still record various types of data. For example, speech recognition has been used to facilitate data entry from humans who are performing various types of inspection procedures which require the use of both hands, such as inventory accounting and cartographic analysis.

2) Speech recognition technology has, in various instances, replaced human operators altogether where procedures are to be initiated upon simple verbal commands. An example of this would be situations where companies receive incoming phone requests for certain types of information, as with stock brokerage firms that typically receive numerous requests for quotations on securities. In such situations, speech recognition technology has proven its ability to eliminate human operators, and to provide required information over phone lines based upon user prompting via alphanumeric input. To this point, only one manufacturer has provided customers with a completely

speaker-independent speech recognition system capable of providing this kind of telephone service.

3) Voice input of data is a highly effective means of obtaining data from humans, as opposed to data entry via keyboard which is a much slower process. Speech recognition technology has also proven its ability to offer a very efficient means of entering commands to computers. A basic illustration of the effectiveness of speech recognition technology involves the fact that humans, too, can more efficiently respond to voice commands, as opposed to visual or cther forms of prompting.

As viewed by the Coast Guard in its Statement of Work for this project, speech recognition technology appears to be most applicable as a means of assisting human operators who monitor distress frequencies. In this sense, speech recognition technology would be intended not so much to replace all human operators, but to provide a low-cost assistance in guarding distress frequencies. At this level speech recognition technology would not actually reduce front-line operating expenses, but would instead be designed to increase the Coast Guard's overall efficiency in monitoring distress signals.

Speech recognition could offer cost savings to back up operators monitoring distress signals. Wherever personnel are now used to back up these operators, speech recognition systems could potentially eliminate those individuals and free them for other duties. Multi-channel recognizers already exist and future generations of recognizers will likely continue this capability. Given this assumption, a single recognition unit could be used to back up several operators through its capability to monitor keywords on several channels simultaneously.

Concluding, speech recognizers do offer definite cost savings advantages in numerous situations. However, in terms of the Coast Guard's application for keyword spotting as a means of backing up human coverage of distress frequencies, its main advantage lies not in terms of its cost effectiveness, but in terms of its overall potential to provide low-cost effective back-up to the Coast Guard's monitoring of distress signals.

6.6 SPEECH SYNTHESIS COST EFFECTIVENESS

This report has already noted that speech synthesis not only saves on manpower required for meeting broadcast requirements, but is also an ideal technology for achieving a more fully automated broadcast facility. As an example of this sort of automation, again refer to the Coast Guard weather reports which are received via teletype. By using

ASCII prompting as input to a voice synthesizer, weather reports could be prepared and stored for broadcast by computer. Weather reports would then consist of stored ASCII input for producing the required broadcasts on a speech synthesizer. This method would entirely eliminate soundproof booths, storage on analog tape, variation in microphone-to-mouth distance, and much of the time it now takes operators to prepare weather reports. Since speech synthesis technology is also an entirely solid-state technology, it should also increase the reliability of the Coast Guard's broadcast system.

The overall degree of cost savings resulting from the use of speech synthesis technology will vary, depending upon the manner in which it is used. For example, many applications have used speech synthesis to aid in setting up efficient man/machine interactions. Other applications have used speech synthesis technology to entirely replace human operators.

6.7 EFFECTIVE USE OF SPEECH SYNTHSIS IN COAST GUARD BROADCASTS

The first and most effective use of speech synthesis would be to broadcast the Coast Guard Weather reports. It is suggested that the Coast Guard might collaborate with the National Weather Service (NWS) in a joint effort to automate this service. This would be in line with a Task Report prepared by the NWS on creating a sample vocabulary of words and/or phrases suitable for use with automated speech generation systems. The NWS previously had evaluated systems capable of providing an automated readout of computer generated weather reports, particularly those used on NOAA Weather Radio. The NWS* had also tested the public's reaction to computer generated forecasts which were broadcasted over WSFC Washington's NWR for a 10-day period. Public reaction to these tests was favorable and indicated that the public would accept a broadcast by a speech synthesizer that basically filled in the blanks of a forecast using pre-recorded phrases selected from a standardized list of permissible expressions.

With the Coast Guard's operational and technical requirements in mind, we suggest that initially analysis synthesis be used for broadcasts to insure the most natural sounding speech. Introducing synthetic speech presents adjustments to the listeners of the broadcasts, so it is essential that that the messages be transmitted with the highest quality speech possible.

We feel that ultimately rule synthesis ha very definite advantages for its potential use. First, rule synthesizers allow an essentially unlimited vocabulary.

* An unpublished working ducument.

Several companies have recently marketed "text-to-speech" synthesizers, where users merely type in the desired phonemes and the unit outputs the desired vocabulary items. We have already mentioned that the Coast Guard receives incoming weather reports via teletype. Since most types of synthesizers accept commands via ASCII, we suggest that the Coast Guard connect a "text-to-speech" rule synthesizer to a teletype, so that incoming weather reports could be prepared for broadcast, as they are received, using synthesis.

Assuming that Coast Guard communications facilities are to be automated, a computer could be used to issue synthesized broadcasts to mariners at specified intervals. This approach eliminates soundproof booths and speaker inconsistencies, since synthesized speech output is uniform and free from background noise and variation in microphone-to-mouth distance. Speech synthesis technology offers the advantage of all solid-state electronics, as opposed to analog recording techniques now used by the Coast Guard. Such electronics have proven to be highly reliable as they contain no moving parts.

Rule-type text-to-speech synthesizers require less programming support than do other types of speech synthesizers; the user merely types in the required text at the keyboard of a CRT terminal, and the synthesizer produces the desired output. Rule synthesizers require little linguistic sophistication on the part of the user. The price of this type of synthesizer has seen a downward trend recently, and should continue to decrease over the next several years. It is anticipated that the quality of speech will improve over time, so that it will be competitive with the very natural sounding analysis synthesis.

Later, when weather reports have been effectively broadcast using speech synthesis, other routine messages could also be automatically generated. Just as there is a suggestion of initially using analysis synthesis for maximally natural sounding speech in weather reports before introducing the extendable rule synthesis, a similar pattern of broadcasting could be done for other messages. The listeners could adjust to more and more messages being broadcast synthetically, if the quality were as humanlike as possible. Then, if unlimited vocabularies were described, rule synthesis which is more machine-like, but capable of generating any and all utterances, including new words and proper names, could be used. By carefully monitoring listener response, the Coast Guard could determine the number and types of messages that should be generated synthetically. Also, records could be kept on the effectiveness of using speech synthesis for Coast Guard broadcasts. It is predicted that the synthetic message will be more and more natural sounding as the technology continues to make advances and that public acceptance of computer generated broadcasts will be regularly increasing.

BIBLIOGRAPHY

Atal, B.S. and Hanauer, S.L. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." Journal of the Acoustical Society of America, Vol.50, No.2, August 1971.

Atal, B. S. and Schroeder, M. "Predictive Coding of Speech Signals." Reports of the Sixth International Congress on Acoustics, ed. by Kohasi, Tokyo, C-5-5, August 21-28, 1968.

Barr, A., Goodnight, J., Sall, J., and Hellwig, J. "A User's Guide to SAS." P.O. Box 10066, Raleigh, N.C.: SAS Institute, Inc., 1979.

Broad, D. "Toward Defining Acoustic Phonetic Equivalence for Vowels." Phonetica. Vol.33, No.6, 1976.

Broad, D., and Fertig, R. "Formant-Frequency Trajectories in Selected CVC Syllable Nucleii." Journal of the Acoustical Society of America. Vol.47, No.6, June 1970.

Coast Guard, "Coast Guard Radio Frequency Plan" (COMDTINST M2400.1 A)

----- "Telecommunications Manual" (COMDTINST M2000.3 A)

Cohen, S., and Pieper, S. "Speakeasy III Reference Manual." 222 W. Adams Street, Chicago, Illinois: Speakeasy Computing Corp. 1979.

Cooper, F. S., Mermelstein, P., and Nye, P.W. "Speech Synthesis as a Tool for the Study of Speech Production." University of Tokyo Press. Haskins Laboratories: Status Report on Speech Research SR-50, 1977.

Doddington, G., and Schalk, T. "Speech Recognition: Turning Theory to Practice." IEEE Spectrum, September 1981.

EDN Magazine, November 20, 1979. "Voice Input and Output."

Gandour, J. "Perceptual Dimensions of Cantonese Tones: A Multidimensional Scaling Reanalysis of Fok's Tone Confusion Data." Sydney, Australia: in SE Asian Linguistic Studies, Vol.4, ed. by Lien, Australian National University Press.

Gandour, J., and Harshman, R. "Crosslanguage Study of Tone Perception." Linguistic Variation: Models and Methods, ed. by Sankoff, New York: New York Academic Press. 1978.

Gray, A., and Markel, J. "Efficient FFT Autocorrelation Estimation." Submitted to IEEE Transactions on Audio and Electroacoustics, September 1972.

Greenberg, J. "Language Universals." The Hague: Mouton, 1966.

Hyman, L. "The Role of Consonant Types in Natural Tonal Assimilations.": in Consonant Types and Tone, ed. by Hyman, USC Occasional Papers in Linguistics, Los Angeles, California. 1973.

----- "Phonology Theory and Analysis." New York, New York: Holt, Rinehart, Winston. 1975.

----- "How Concrete is Phonology?" Language, 46.58-76, 1970.

Ingemann, F. "Speech Synthesis by Rule Using the FOVE Program." Paper presented at the IPS-77, Miami Beach. Haskins Laboratories: Status Report on Speech Research SR-54,1978.

Itakura, F., and Saito, S. "Digital Filtering Techniques for Speech Analysis and Synthesis." 7th. International Congress on Acoustics, Budapest, paper 25 C 1, 1971.

Jakobson, R., Fant, G., and Halle, M. "Preliminaries to Speech Analysis The Distinctive Features and their Correlates." The MIT Press, Massachusetts Inst. of Technology, Cambridge, Mass., 1972.

Kelly, R. E. "Speech and Hearing: Using them to Assist Offshore Systems Operators." Southwest Research Institute. Paper presented at the 11th. Annual OTC in Houston, Texas. April 30-May 3, 1979.

Klatt, D. H. "Review of the ARPA Speech Understanding Project," Journal of Acoustical Society of America, Vol.62, No.6, December 1977.

Klatt, D. "Software for a Cascade/Parallel Formant Synthesizer." Journal of the Acoustical Society of America. Vol.67, No.3, March 1980.

Ladefoged, P. "Preliminaries to Linguistic Phonetics." Chicago: University of Chicago Press, 1971.

Lea, W. A. (ed.) "Trends in Speech Recognition," Englewood Cliffs, N. J.: Prentice-Hall, Inc.1980.

Lehiste, I. "Influence of Fo Pattern on the Perception of Duration." Acoustical Society of America Paper, 1975.

----- "Suprasegmentals." Cambridge, Mass.: MIT Press, 1970.

Makhoul, J., Viswanathan, R., and Huggins, W. F. "A Mixed-Source Model for Speech Compression and Synthesis." Bolt, Beranek, and Newman Incorporated, Cambridge, Massachusetts 02138. Journal of Acoustical Society of America, Vol.64, No.6, December 1978.

Markel, G., and Gray, A. "On Autocorrelation Equations with Application Fo Speech Analysis." IEEE Transactions on Audio and Electroacoustics, Vol.Au-21, No.3, 1973.

Markel, G. "Basic Formant and Fo Parameter Extraction from a Digital Inverse Filter Formulation." IEEE Transactions on Audio and Electroacoustics. Vol.Au-21, No.3, 1973.

Mattingly, I. G. "Phonetic Representation and Speech Synthesis by Rule." Paper presented at the International Symposium on the Cognitive Representation of Speech, Edinburgh, July 29-Aug 1, 1979. Haskins Laboratories: Status Report on Speech Research SR-61, 1980

May, J. C. "Speech synthesis Using Allophones," Speech Technology, Vol.1, No.2, April 1982, pp. 58-62.

Mermelstein, P. and Rubin, P. "Articulatory Synthesis_ A Tool for the Perceptual Evaluation of Articulatory Gestures." Paper presented at the Symposium on Articulatory Modelling, Grenoble, France, 11-12 July, 1977. Haskins Laboratories: Status Report on Speech Research SR-53, Vol. 1, 1978.

Nielsen, A. S. "Listener Preference and Comprehension Tests of Stress Algorithms for a Text-to-phonetics Speech Synthesis Program." Naval Research Laboratory, Washington D.C. 1976.

Noll, A. "Cepstrum Pitch Determination." Journal of the Acoustical Society of America, Vol.41, No.2, February 1967.

Noll, A. "Short-time Spectrum and Cepstrum." Techniques for Vocal-Pitch Detection." Journal of the Acoustical Society of America, Vol.36, No.2, February 1964.

O'Shaughnessey, D. "The Search for Fo Features." Cambridge, Mass.: Massachusetts Institute of Technology, 1977.

Ohala, J. "The Physiology of Tone." in Consonant Types and Tone, ed. by Hyman, USC, Los Angeles, 1973.

Pisoni, D. B. "Perception of Speech: The Human Listener as a Cognitive Interface," Speech Technology, Vol.1, No.2, April 1982, pp. 10-24.

Purcell, E., and Suter, R. "Predictors of Pronunciation Accuracy." Language Learning, 1978.

Rabiner, L., Sambur, M., and Schmidt, C. "Applications of a Nonlinear Smoothing Algorithm to Speech Processing." IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol.ASSP-23, No.6, December 1975.

Roberson, J. "Correlation Between Fo and Vowel Duration in English: a Statistically Oriented Pretest." USC Department of Linguistics, Los Angeles, 1978.

----- "Path Analysis Model for Word Order Universals." Los Angeles, California: USC Department of Linguistics, unpublished, 1979.

----- "Speech Recognizer Standard Testing and Variables Which Affect Speech Recognizer Performance." SCRL paper, unpublished, 25 July, 1980.

Ryan, T., Joiner, B., and Ryan, B. "Minitab II Reference Manual" N. Scituate, Mass. 02060: Dunsbury Press, 1976.

Schafer, R., and Rabiner L. "System for Automatic Formant Analysis of Voiced Speech." Journal of the Acoustical Society of America. Vol.47, No.2, 1970.

Coast Guard, "Telecommunication Mannual" (COMDTINST M2000.3A)

Vetter, D., Stork, J., Skoge, K., and Ahrens, P. "LPC Speech I. C. using a 12-pole cascade digital filter." Reprinted from proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, March 30,31 and April 1, 1981.

Wiggins, R. "Development and Application of LSI Speech Synthesizers." Texas Instruments, Inc. Dallas, Texas, 1980.

Wiggins, R., and Brantingham, L. "Three-Chip System Synthesizes Human Speech." Electronics Magazine, August 31, 1978.

Zee, E "Duration and Intensity as Correlates of Fo." Acoustical Society of America conference paper, 1976.

Zink, H.E., and Plummer, W. "Automated Voice Readout of Computer Generated Weather Reports." National Weather Service funded under Task 7620 to Navy Contract N00017-72-C-4401. The Johns Hopkins University Applied Physics Laboratory. Johns Hopkins Road, Laurel, Maryland 20810. Speech Technology, Vol.I, No.2, April, 1982.

APPENDIX A

GLOSSARY OF TERMS USED IN THIS REPORT

This glossary is intended to present to interested readers definitions of various terms used in this report. These definitions are designed to facilitate comprehension of this report by those whose technical expertise may lie outside the area of voice technology.

1) *accuracy rate* - performance measures given by manufacturers of their recognition systems. These figures must be regarded with some care. The accuracy rates are based on different types of vocabularies, since there is not any widely accepted vocabulary used to test recognizers. Each manufacturer is generally free to use its own chosen vocabulary. Naturally, some vocabularies are easier than others for recognition success. For example, "right" vs. "left" and "up" vs. "down" are easier to distinguish acoustically from one another than are "right" vs. "ripe" and "down" vs. "done". Typically, manufacturers choose sample recognition vocabularies that are maximally effective for their own recognizers. Thus, we

notice that rarely will any manufacturer of speech recognizers claim an accuracy rate of less than 95 - 99%. Thus, it is extremely important to test recognizers with the intended vocabulary for the user who will purchase the system in order to obtain a better idea as to what the accuracy rate of the recognizer will actually be in the field.

2) adjustable reject setting - variation of the acceptance threshold for vocabulary items. For example, if the adjustable reject setting is too high, the accuracy rate will be reduced, since only input vocabulary items with a very high degree of statistical correspondence with stored templates will be recognized. On the other hand, if the reject level is set too low, false recognitions can be generated, since there is a relatively lower criterion for matching input vocabulary with stored templates. Part of the solution to this problem involves the careful choice of vocabulary items, so that they are maximally distinct acoustically.

3) analysis synthesis - a type of speech generation or speech synthesis that is based upon an acoustic analysis of real human speech. Basically, an analysis of human speech can be used to define the

gross values for a digital filter simulating human speech. This type of synthesis is distinguished from rule synthesis which does not base output speech upon real human speech, but on basic combinations of acoustic parameters which produce human-like speech. A common type of speech analysis used for analysis synthesis is linear prediction coding known as LPC synthesis.

4) bit rate - the number of bits that are used to synthesize digitally a speech utterance. The lower the bit rate, the less information that has to be stored in computer memory. This is an important consideration for cost effective speech synthesizers. The higher the bit rate, the more information that contributes to natural sounding speech synthesis.

5) byte - one unit of information in computing. On IBM systems, there are 8 bits per byte. On ASCII terminals, there are 7 bits per byte.

6) coarticulations - the changes in the acoustic parameters of speech which occur between adjacent vowel and consonant sounds. When individual speech sounds are joined together to form words and sentences, certain of their acoustic parameters are affected by the neighboring sounds

or phonemes. Accounting for all possible coarticulations between phonemes is a difficult procedure, yet one that is essential for producing the best speech synthesis.

7) codec - a device which stores speech data which have been digitally encoded.

8) computer storage - both "real storage" which is the amount of storage required in the central processing unit of the computer and "virtual storage" which is storage within the computer, but not part of the central processing unit.

9) connected or continuous speech recognizer - a recognizer that is able to correctly identify input speech which consists of concatenated, or connected, strings of words. This is a much more difficult task than recognition by isolated word recognizers, for the words flow continuously together and do not have boundaries of silence between them.

10) EPROM - eraseable, programmable, read only memory category of LSI (large scale integrated) chips.

11) formant - an acoustic resonant frequency and its associated bandwidth. Every speech sound has a set of formants which are determined by the configuration of the vocal tract.

12) fricative speech sounds - productions of speech which are predominantly turbulent, since they have a noise source at the place of articulation. Examples of fricatives include the first sound in each of the following words; fun, very, he.

13) isolated word recognizer - a recognizer that is able to handle only single, or isolated, words which are not embedded in phrases or sentences, but are pronounced with boundaries of silence surrounding them. This is an easier recognition task than that of handling "connected speech" which consists of words strung together to form whole sentences, at a regular repetition rate. Isolated word recognizers generally perform best where input vocabulary items are pronounced relatively slowly and precisely, both during training and actual recognition.

14) keyword spotting - recognizing certain words of specific interest in connected speech input. The Coast Guard has seen keyword spotting of such words as "mayday" or "fire" as a means of backing up human operators who are assigned to monitor radio receptions on distress frequencies.

15) LPC synthesis - a type of speech generation or speech synthesis that uses a "linear prediction

coding" model of speech. This approach of analysis synthesis uses a digital filter to model the human vocal tract; it is based upon the statistical assumption that human speech changes relatively slowly, and that it is possible to predict the next set of acoustic measures based on a knowledge of previous ones.

16) LSI chip - large scale integrated circuits which are put into a single chip. The LSI chip industry is a key part of speech technology. As the price for such chips has decreased through volume production methods, integrating speech technology into new product areas has become more attractive.

17) nasal speech sounds - productions of speech which are made with the air stream being emitted only through the nose. Examples of nasals include the first sound in the word mat and in the word nice.

18) nonperiodic sound - a sound which does not have a waveform with a consistently repetitive rate. For example, vowels are characterized by having waveforms which are basically periodic in nature, but plosives and fricatives do not have such waveforms. These consonants have nonperiodic waveforms of burst and turbulence. We have noticed that nonperiodic sounds were commonly

found in Coast Guard radio receptions, such as pops and clicks. These nonperiodic sounds can sometimes generate false recognitions when mistakenly identified as plosives or fricatives.

19) O.E.M. - original equipment manufacturers. This includes companies that integrate commercially purchased items, such as LSI chips, into their own products which are again sold as a finished product.

20) performance analysis - a statistical analysis of human or human/equipment performance. The basic concept involves identifying variables which significantly affect human or human/equipment performance of a predefined task through multiple regression and factor analysis. The term performance analysis refers to actually deriving a regression equation which describes human or human/equipment interactions. Such an equation would be helpful by identifying variables which significantly affect recognition accuracy rates. If we actually had a performance analysis of speech recognizers which gave significant results, we might be able to identify which recognition algorithms were relatively preferable to others.

21) phonemes - basic units of sound in human speech, sometimes referred to as the vowels and consonants of the language. Phonemes are discrete sounds which can cause a difference in meaning between otherwise identical words, such as "bat" and "pat" or "but" and "bit".

22) phonetic context - the location of an acoustic entity with reference to surrounding sounds. The acoustic context has a noticeable effect upon phonemes. Linguists commonly refer to allophones which are pronunciation variants of basic phonemes due in part to their phonetic contexts. Since a given phoneme can have a variety of allophones, this makes the overall recognition task more difficult, particularly when attempting to develop a speaker-independent recognizer which will handle connected speech wit' various allophones.

23) plosive speech sounds - productions of speech which involve a blocking or stoppage of the air flow from the vocal tract. Examples of plosives include the first sound in each of the following words: pal, door, give.

24) prosodics - the "suprasegmentals" or influences of duration, fundamental frequency, and speech production power upon basic phonemes. These

influences include emphasis, stress, and durational patterns of the vowels and consonants. The prosodic parameters are basic to both the recognition and synthesis of speech.

25) rapid speech - speech which is produced considerably faster than carefully articulated speech.

26) ROM - read only memory LSI chip category, not erasable or programmable.

27) rule synthesis - a type of speech generation or speech synthesis that uses a set of rules to model speech. This approach specifies which combinations of acoustic parameters are to be used to best imitate human speech.

28) sampling rate - the frequency with which speech recognizers digitally encode speech data. Such digitized data are actually numerical codings which represent the translation or analysis of the real speech waves. The numerical data are taken at uniform points along these speech waves and expressed as a function of time. A common sampling rate for laboratory analysis of speech waves is 10,000 samples per second. Commercial recognizers, however, tend toward a lower sampling rate to conserve memory.

29) speaker dependent recognizer - a recognizer that requires "training" by an individual before recognition of that individual's speech can take place. "Training" generally consists of having the person whose vocabulary is to be input for recognition repeat this vocabulary several times, so that templates can be established for each vocabulary item of a given individual. Such templates are then used for comparison with incoming vocabulary items.

30) speaker independent recognizer - a recognizer that does not require "training" by individual speakers before recognition of that individual's speech can take place. A speaker independent recognizer allows virtually any person to input speech with no stored information about that speaker's characteristics to aid the machine in its recognition of the speech.

31) speech recognizer - a device that accepts speech as an input and produces typed messages or action by a machine controlled by the recognizer as output. A speech recognizer can be digital or analog. It can be one of two types: 1) isolated word, or 2) connected or continuous speech.

32) speech synthesizer - a device that generates speech mechanically. A speech synthesizer can be digital or analog. It can be one of three types: 1) analysis synthesizer, 2) rule synthesizer, or 3) digital recoding synthesizer.

33) templates - acoustic manifestations of words or utterances that have been stored in digital form. Templates are actually sets of numbers representing acoustically derived parameters. When training speech recognizers, templates are set up as a speaker repeats his/her vocabulary items into the recognizer. Later, these templates are used as references for statistical comparison to input vocabulary items to determine the identity of the input vocabulary.

34) voice recognition - either automatic recognition of words and sentences which are spoken into a speech recognizer or automatic recognition of the voice quality of the speaker, thus serving as an identification of the person who is speaking. In the first case, "voice recognition" is synonymous with "speech recognition" and in the second case with "speaker recognition"

35) voiced speech sounds - productions of speech involving a vibration of the true vocal folds.

Examples of voiced sounds include the first sound in each of the following words: it, me, big.

36) voiceless speech sounds - productions of speech not involving a vibration of the true vocal folds. Examples of voiceless sounds include the first sound in each of the following words: keep, saw.

APPENDIX B

DESCRIPTION OF ILS

The ILS commands have been written to utilize peripheral sevices such as disk packs for data storage and retrieval, the line printer for listings, and the analog-to-digital and digital-to-analog converters as means of interfacing the analog representation of signals with the digital representation. The means of interaction with the system is a terminal which has graphic as well as alphanumeric (text) capabilities.

The ILS software has been developed as a set of self-contained program modules which are utilized serially. Each ILS command is a program module which executes a specific task. The program modules are stored on disk and are brought into core one at a time by user command. Consequently, except for the keyboard monitoring program, the memory resources of ILS are only in demand when an ILS command is being executed.

Thus, ILS is not taking up memory space during the time the user is just thinking, examining a display, toying the next command, etc. This is an important factor in multi-user systems which may have memory limitations.

The critical provisions for communication of parameter values between program modules is made possible by providing on disk an exclusive file for each user. This file, conveniently designated as the user's COMMON file, contains global system parameters and it serves as a work area for deposit and retrieval of information by all commands executed by the user. The acquisition of shared information is affected as each module initiates its execution by reading the disk-resident user's COMMON file. At the conclusion of its execution each module then rewrites back onto disk the updated version of the user's COMMON file. In this way an TLS module can operate on previous results and arguments passed through the user's COMMON file by a preceding module.

Because of the modularity of the system, any program module may be modified without affecting the other modules. This feature also permits the replacement or addition of program modules on disk providing they are properly designed to be compatible with the ILS conventions. Thus, each user can have his own tailored ILS commands.

Principles of Operation

The software complement of the Interactive Laboratory System has been designed to function entirely under the control of the computer's own operating system. In this way the ILS modules take advantage of existing subroutines and file structures available within the computer processing system.

It may be helpful at the outset to describe the memory organization of the computer system very simply as having two working segments. One segment is occupied by the computer operating system (at all times) and the other is available for the execution of programs entered by various computer users. The computer operating system can be described in most general terms as an operating executive which allocates computer resources to a set of users. This system itself consists of a collection of programs and tables which are used to control the flow of information processing within the computer.

The Interactive Laboratory System is an organized collection of interrelated but independent program modules. These disk-resident modules are independent in the sense that each module becomes the sole occupant of the user segment of core when called out by user command, and thus renders a solo performance as far as the remaining disk-bound modules are concerned. The interrelatedness of the ILS program modules is realized through the passing of constants, variables, and arrays through each user's COMMON file from one successive module to another.

It may further be helpful to identify the actual nature of the program modules as they are placed on disk. The modules actually consist of files of binary data which are computer translations into machine language of the original FORTRAN source program written by the ILS programmers. When read into core, these files become operating intelligence in executing the objectives written into a module. In order to do this, the processing system of the computer in effect places itself at the disposal of the ILS program currently resident within the core and implements its instructions.

APPENDIX C

LIST OF MANUFACTURERS

Manufacturers of Speech Recognition Products:

Centigram Corp. 155A Moffett Park Drive, Suite 108, Sunnyvale, California 94086 (408) 734-3222

Heuristics, Inc. 1285 Hammerwood Avenue, Sunnyvale, California 94086 (408) 734-8532

Interstate Electronics Corp. (ATO subsidiary) 1001 E. Ball Road, P.O. Box 3117, Anaheim, California 92805 (714) 635-7210

Nippon Electric Company, Ltd. NEC America, Inc. 532 Broadhollow Road, Melville, New York 11746 (516) 752-9700

Scott Instruments 815 North Elm, Denton, Texas 71201 (817) 387-9514

Threshold Technology, Inc. 1829 Underwood Place, Delran, New Jersey 08075 (609) 461-9200

Verbex Corp. (Exxon subsidiary) 2 Oak Park, Bedford, Massachusetts 01730 (617) 275-5160

Voicetek P.O. Box 388, Goleta, California 93017 (805) 685-1854

Votan, Inc. 26046 Eden Landing Road, Suite 7, Hayward, California 94545 (415) 785-8060

Manufacturers of Speech Synthesis Products:

Centigram Corp. 155A Moffett Park Drive, Suite 108, Sunnyvale, California 94086 (408) 734-3222

General Instruments Corp. Microelectronics Division, 600 West John Street, Hicksville, New York 11802 (516) 733-3107

Interstate Electronics Corp.(ATO subsidiary) 1001 E. Ball Poad, P.O. Box 3117, Anaheim, California 92805 (714) 635-7210

Kurzweil Computer Products, Inc. 33 Cambridge Parkway, Cambridge, Massachusetts 02142 (617) 864-4700

Maryland Computer Services, Inc. 502 Rock Spring Avenue, Bel Air, Maryland 21014 (301) 838-8888

Mimic Electronics P.O. Box 921, Acton, Massachusetts
01720 (617) 263-2101

MSC. 1640 Monrovia, Costa Mesa, Ca. 92627 (714)
642-2427

National Semiconductor Corp. 2900 Semiconductor Drive,
Santa Clara, California 95051 (408) 737-5864

Percom Data Co., Inc. 211 North Kirby, Garland, Texas
75042 (214) 272-3421

Telesensory Speech Systems, Inc. 3408 Hillview Avenue,
Palo Alto, California 94304 (415) 493-2626

Texas Instruments 8600 Commerce Park Drive, Suite [illegible],
Houston, Texas 77036 (713) 776-6511

Votrax 500 Stephenson Highway, Troy, Michigan [illegible].
(313) 588-2050

ATE

LMED

8